INSTITUTO SUPERIOR DE ESTATÍSTICA E GESTÃO DA INFORMAÇÃO
DA
UNIVERSIDADE NOVA DE LISBOA

A APLICAÇÃO DE REDES NEURONAIS NA DETECÇÃO DA INFLUÊNCIA DO HIGH FREQUENCY TRADING NA NEGOCIAÇÃO DE ACÇÕES (CASO PORTUGUÊS)

Wellington Ferreira de Oliveira

Lisboa, 2011

INSTITUTO SUPERIOR DE ESTATÍSTICA E GESTÃO DA INFORMAÇÃO
DA
UNIVERSIDADE NOVA DE LISBOA

A APLICAÇÃO DE REDES NEURONAIS NA DETECÇÃO DA INFLUÊNCIA DO HIGH FREQUENCY TRADING NA NEGOCIAÇÃO DE ACÇÕES (CASO PORTUGUÊS)

Wellington Ferreira de Oliveira

Versão definitiva da Proposta de Dissertação a apresentar como requisito parcial para obtenção do grau de Mestre em Estatística e Gestão da Informação

Professor orientador:
Professor Doutor Fernando Bação

Lisboa, 30 de Novembro, 2011

Agradeço ao meu orientador, Professor Doutor Fernando Bação, pela sua disponibilidade, sapiência e orientação na concretização deste trabalho. A professora Sandra Catarino, pela sua sempre bem-disposta disponibilidade.

Aos meus colegas e amigos pelo apoio e entusiasmo, não só para realização mas, principalmente para concretização deste trabalho. A Vossa paciência e colaboração, nas revisões e críticas, foram fundamentais.

À Euronext Lisbon, nomeadamente ao Dr. Luis Laginha, por ter autorizado a utilização dos dados que serviram de base para esta dissertação.

E principalmente à minha família. À Mônica, pelo apoio incondicional e incentivo à conclusão deste trabalho. À Beatriz, por ser o meu incentivo em tentar fazer todos os dias algo melhor.

Muitíssimo Obrigado a todos.

RESUMO

O aumento da capacidade de processamento dos sistemas de negociação, em grande parte reflexo do desenvolvimento tecnológico verificado na última década, transformou a negociação no mercado de instrumentos financeiros, sendo que actualmente volumes gigantescos de dados de negociação e o *high frequency trading* (HFT) são fenómenos indissociáveis.

Nesta nova realidade, onde decisões de investimento passam a estar suportadas em algoritmos electrónicos em detrimento da acção humana, a indústria, as sociedades gestoras de sistema e plataformas e os reguladores europeus, discutem a necessidade de estabelecer, ou não, limites à utilização do HFT.

Neste confronto entre vantagens e desvantagens, entre o permitir e o regular, é necessário conciliar a liquidez, a regularidade de funcionamento e a confiança dos investidores num meio ambiente envolvente que funciona à velocidade da luz, onde o controlo do risco e a detecção de situações de abuso de mercado colocam novos desafios a todos os participantes no mercado. As recomendações da Organização Internacional das Comissões de Valores (IOSCO) publicadas em Outubro de 2011 traduzem de forma inequívoca esta preocupação.

O presente trabalho utiliza uma rede neuronal artificial não supervisionada para, pela primeira vez, detectar a significância do fenómeno de HFT no mercado de acções nacional.

Recorrendo a um Self-Organizing Map (SOM) e utilizando dados da negociação referentes a ofertas e sobre uma amostra de acções que integram o índice PSI20 foi possível distinguir os intermediários financeiros com comportamento característico de um *High Frequency Trader* (HFTr).

Palavras-chave: *High Frequency Trading, High Frequency Trader*, Redes neuronais, *Self-Organizing Map*, Segmentação, Euronext Lisbon.

ABSTRACT

The increasing in the processing capacity of trading systems, largely as a reflection of technological developments over the last decade, has transformed trading in financial instruments markets, as it, nowadays, massive volumes of transaction data and high frequency trading (HFT) are strictly linked.

Within this new reality, where investment decisions are now supported by electronic algorithms rather than human decisions, market industry and operators or investment firm managed of multilateral systems, as well as European regulators, have been intensively discussing the need to establish (or not), limitations to the use of HFT techniques.

The confrontation between advantages and disadvantages arising from the free permission of HFT or from regulating it, requires one to consider the impacts on liquidity, on the regularity of market functioning, as well as over investors´ confidence in a surrounding environment that 'flows at the speed of light'. One must also consider the new challenges posed to all market participants in what respects to risk control and detection of market abuse. The International Organization of Securities Commissions (IOSCO) recommendations, published in October 2011, unequivocally demonstrate this concern.

This paper uses an unsupervised neural network, employed, as long as we know, for the first time to detect the significance of HFT in the Portuguese stock market.

Using a Self-Organizing Map (SOM) and data relating to offers and trading on a sample of shares that integrate the PSI 20 index, it was found evidence that some financial intermediaries act in the market in a way similar to High Frequency Traders' (HFTr) characteristics and behavior.

Keywords: High Frequency Trading, High Frequency Trader, Neural networks, Self-Organizing Map, classificação, Euronext Lisbon.

ÍNDICE

SIGLAS E ABREVIATURAS

AA	Algoritmos de agência
AFM	Authority for the Financial Markets (Regulador Holandês)
AP	Algoritmos proprietários
AT	Algoritmo electrónico de negociação (*Algorithm trading*)
CE	Comissão Europeia (*European Commission*)
CESR	*The Committee of European Securities Regulation* (actual ESMA)
CMVM	Comissão do Mercado de Valores Mobiliários
DMA	Acesso Directo ao Mercado (*Direct Market Access*)
DMIF	Directiva de Mercado e Instrumentos Financeiros (*Mifid*)
ESMA	*European Securities and Markets Authority* (*ex* CERS)
ET	Sistemas de Negociação electrónicos (*Electronic Trading*)
HFT	Negociação de Alta frequência (*high-frequency trading*)
HFTr	*High Frequency Trader* (IF que utiliza a solução HFT)
HHI	*Herfindahl-Hirschman Index*
IF	Intermediários Financeiros
IOSCO	Comitê Técnico da Organização Internacional das Comissões de Valores
ISIN	Código de Identificação do instrumento financeiro (*International Securities Identification Number* – ISO 6166)
LIST	Lisbon Trading
MM	IF que atribui liquidez ao mercado (*Market Maker*)
MTF	Sistemas de Negociação Multilateral (*Multilateral Trading Facilities*)
NSC	Sistema de negociação da EURONEXT (*Nouvelle Système de Cotation* – até 2009)
PT	Programas electrónicos de negociação (*Program trading*)
RNA	Redes Neuronais artificiais
SA	Acesso patrocinado (*Sponsored Access*)
SOM	*Self Organizing Map*
TWAP	*Time weighted average price* (estratégia de negociação)
UTP	Sistema de negociação da EURONEXT (*Universal Trading Platform* desde 2009)

1. INTRODUÇÃO

A harmonização das estruturas de mercado, a adopção de um enquadramento jurídico destinado a regular a concorrência entre os mercados regulamentados[1] europeus e a criação de duas novas formas organizadas de negociação, os sistemas de negociação multilateral[2] e a internalização sistemática[3], introduzidas pela Directiva de Mercado e Instrumentos Financeiros (DMIF)[4], intensificaram a concorrência entre sociedades gestoras de sistemas de negociação multilateral, historicamente conhecidas como *Bolsas de Valores*.

É neste cenário de maior flexibilidade de negociação e consequente evolução tecnológica, onde decisões de investimento passam também a estar suportadas em algoritmos electrónicos em detrimento da acção humana, que o tema *high-frequency trading* (HFT)[5] tem ganho relevo. Mais do que um mero subproduto deste novo cenário competitivo, o HFT está a mudar a forma de actuação dos Intermediários Financeiros (IF) na negociação em bolsa.

---

[1] *Mercado regulamentado* - sistemas multilaterais, autorizados como tal num qualquer Estado membro da UE, de funcionamento regular e que possibilitam o encontro de interesses relativos a instrumentos financeiros com vista à celebração de contratos sobre tais instrumentos – art. 4/1/14 DMIF, p. 10.

[2] MTF – *Multilateral Trading Facilities* ou Sistema de Negociação Multilateral estão autorizados num qualquer Estado membro da EU, e possibilitam o encontro de interesses relativos a instrumentos financeiros com vista à celebração de contratos sobre tais instrumentos - art. 4/1/15 DMIF, p.10.

[3] SI – *Systematic Internalization* ou Internalização Sistemática é a negociação por conta própria, em execução de ordens de clientes, por intermediário financeiro, fora de mercado regulamentado ou de sistema de negociação multilateral, de modo organizado, frequente e sistemático – art. 4.º/1/7 DMIF, p.10.

[4] A DMIF é uma directiva europeia (Directiva 2004/39/CE do Parlamento Europeu e do Conselho de 21 de Abril de 2004, que foi aplicada pela Directiva 2006/73/CE da Comissão de 10 de Agosto de 2006) que visa uma maior harmonização da legislação europeia e tem como objectivo principal a criação de um mercado único europeu de serviços financeiros, baseado numa maior transparência na negociação de um vasto leque de instrumentos financeiros e no aumento da protecção dada ao investidor, ajustando-a às suas características, experiência e conhecimentos financeiros. Esta Directiva foi transposta para a legislação nacional e entrou em vigor a 1 de Novembro e 2007.

[5] Por uma questão de facilidade e equiparação às práticas de mercado optamos por manter, a par do anglicismo, a terminologia utilizada no mercado financeiro, representada pela sigla HFT, para identificar um conjunto específico de operações realizadas em condições especiais. Na tradução portuguesa, poder-se-ia identificar como "Negociação de Alta Frequência".

O HFT, actualmente intensamente discutido no contexto europeu[6], já há bastante tempo constitui um tema central no funcionamento dos mercados financeiros desenvolvidos, com especial relevância no mercado norte-americano (EUA), onde no auge da crise financeira de 2008, diversos investidores obtiveram elevados ganhos[7] através da optimização de estratégias de negociação. Em estudo recente, encomendado pelo governo Inglês, o tema adquire relevo ao referir que um terço do volume de negociação de acções realizado no Reino Unido é gerado através de HFT, enquanto nos EUA, este número já chega a três quartos[8].

Os novos sistemas multilaterais de negociação (MTF), a redução do tempo de latência das estruturas de negociação[9], o incremento do número de transacções diárias, a redução da quantidade média de instrumentos financeiros[10] negociados por operação[11], as alterações na estrutura de mercado decorrentes do aumento do nível de liquidez, a redução da volatilidade[12] e do *spread* [13], contribuem para um ambiente propício à implementação de estratégias de investimento suportadas em algoritmos electrónicos de negociação (AT)[14] , e, simultaneamente como um resultado dessas mesmas estratégias.

Por outro lado, a ocorrência de erros nos programas automáticos de envio de ofertas para os sistemas de negociação e a fragilidade de alguns IF em adaptarem as suas estruturas de negociação ao significativo aumento do número de mensagens (ofertas) diárias, têm dividido opiniões entre os especialistas de mercado, as entidades

---

[6] Tema constante de discussões e consultas realizadas pelo Colégio de Reguladores (CESR, actual ESMA) e pela Comissão Europeia no âmbito dos trabalhos em curso para a revisão da DMIF.

[7] Segundo Irene Aldridge (2010, p.1), mesmo nos piores meses da crise de 2008, mais de 50% das operações realizadas no mercado financeiro (de bolsa) envolviam HFT.

[8] *The Future of Computer Trading in Financial Markets* (2011). Retrieved October 06, 2011 from: http://www.bis.gov.uk/assets/bispartners/foresight/docs/computer-trading/11-1276-the-future-of-computer-trading-in-financial-markets .

[9] Ver o conceito de *latência* apresentado no capítulo 7.

[10] Instrumento financeiro – "qualquer dos instrumentos especificados na Secção C do Anexo I" (valores mobiliários e outros nove tipos diferentes de instrumentos) - art. 4/1/17 DMIF, p.10.

[11] Entenda-se "operação" o resultado do encontro das intenções de compra e venda de um determinado instrumento financeiro e ao mesmo preço. Neste estudo resultado do encontro de ofertas sobre acções.

[12] Ver conceito de *volatilidade* apresentado no capítulo 7.

[13] O *spread* é a terminologia financeira utilizada para identificar a diferença entre o melhor preço de venda e o melhor preço de compra disponível a cada momento.

[14] *Algorithm trading* (algoritmos electrónicos de negociação) é a terminologia inglesa utilizada para identificar programas automáticos de envio de ofertas para os sistemas de negociação. Suportados em modelos matemáticos constituem hoje uma solução de negociação não só no mercado de valores mobiliários (acções), como também no mercado de derivados (futuros e opções).

gestoras de sistemas de negociação multilateral (Bolsas de Valores)[15] e os reguladores[16] quanto ao estabelecimento, ou não, de limites à utilização do HFT.

A par destas preocupações o Comitê Técnico da Organização Internacional das Comissões de Valores (IOSCO) publicou em Outubro de 2011[17], o relatório final de uma consulta pública, sobre o impacto da evolução tecnológica na integridade e na eficiência do mercado de instrumentos financeiros. Neste relatório a IOSCO analisa os desenvolvimentos tecnológicos mais relevantes verificados nos mercados financeiros nos últimos anos, e estabelece recomendações aos reguladores para implementação de procedimentos à mitigação dos possíveis riscos relacionados com os algoritmos automáticos de negociação e o HFT.

Segundo Beddington (2011), "embora a prevalência da negociação electrónica seja uma realidade, encontramo-nos diante de diversos pontos de vista diferentes sobre os riscos e os benefícios em que esta se traduz", bem como o que poderá produzir no desenvolvimento do mercado financeiro mundial. Neste sentido, "obter uma melhor compreensão destas questões é fundamental" (Beddington, 2011, p.2).

Compreender então, em que medida, este fenómeno faz já parte da realidade portuguesa é imprescindível, quer para uma adequada supervisão da negociação, quer para a mensuração dos possíveis riscos envolvidos.

1.1. O Contexto do Mercado Português.

Quando nos referimos ao contexto estrito de actuação dos profissionais de mercados (*Traders*), aquando da satisfação do interesse dos seus clientes (*investidores*) junto de um sistema multilateral de negociação, verificamos que há pouco mais de uma década o mercado de valores mobiliários[18] português reconhecia uma nova etapa da

---

[15] Por uma questão de facilidade de interpretação utilizaremos a terminologia Sociedades Gestoras de Mercado para definir "*Operador de mercado*: a pessoa ou pessoas que gerem e/ou operam as actividades de um mercado regulamentado. O operador de mercado pode ser o próprio mercado regulamentado" – art. 4.º/1/13 DMIF, p.10.

[16] Em Portugal a Comissão do Mercado de Valores Mobiliários (CMVM).

[17] *Regulatory Issues Raised by the Impact of Technological Changes on Market Integrity and Efficiency*. Retrieved November 06, 2011 from: http://www.iosco.org/library/pubdocs/pdf/IOSCOPD361.pdf

[18] "Valores mobiliários - categorias de valores que são negociáveis no mercado de capitais, com excepção dos meios de pagamento, como por exemplo: a) Acções de sociedades e outros valores

evolução tecnológica ao serviço da negociação em bolsa de valores. A introdução de um novo sistema electrónico de negociação na então Bolsa de Valores de Lisboa[19] (1999), o LIST – *Lisbon Trading*[20], veio permitir aos IF portugueses[21] uma nova forma e uma melhor qualidade de execução das intenções de compra e venda (ofertas)[22] dos seus clientes, verificando-se desde logo um significativo incremento do número de ofertas diariamente transmitidas ao sistema de negociação e na quantidade de negócios realizados[23] em cada sessão de bolsa. O novo sistema de negociação introduziu ainda uma nova dinâmica na gestão das intenções de investimentos (ofertas) por parte dos IF, permitindo registar, quer antes do início da negociação[24], quer durante a sessão de bolsa em contínuo, num curto intervalo de tempo, um conjunto significativo de ofertas previamente configuradas[25] nas suas estações de trabalho (*front-end*), o que abriu um novo leque de oportunidades na programação atempada de estratégias de investimento. Outra novidade indroduzida com o LIST foi a implementação de mecanismos ligados em rede, identificados como "*order collecting*"[26].

---

equivalentes a acções de sociedades, de sociedades de responsabilidade ilimitada (*partnership*) ou de outras entidades, bem como certificados de depósito de acções; ……" - 4.º/1/18 DMIF, p.10.

[19] Em Dezembro de 1999, resultante de um novo enquadramento regulamentar, as duas Associações de Bolsa (Lisboa e Porto) foram transformadas numa única sociedade anónima que integrou a gestão do mercado de operações a contado e a prazo, a Bolsa de Valores de Lisboa e Porto (BVLP).

[20] O LIST entrou em funcionamento em 1 de Março de 1999, coincidindo com a entrada em vigor do Regulamento Nº1/99 da CMVM, que estabeleceu novas regras de negociação para o mercado a contado. O *benchmarketing* do LIST era a garantia do processamento de 98% das ofertas em menos de 2 segundos.

[21] Membros Negociadores – designação posteriormente introduzida com o novo sistema de negociação.

[22] Por oferta (*order* na terminologia inglesa), entenda-se a intenção de compra ou venda registada numa qualquer forma organizada de negociação.

[23] O número de ofertas inseridas no sistema de negociação (LIST) em Março de 1999 foi de cerca de 25.000. No entanto o número de negócios no primeiro ano de funcionamento (1999) manteve-se equivalente ao realizado em 1998, cerca de 2,3 milhões. Em 1997 o número total de negócios realizados na BVL foi cerca de 860.000.

[24] No período designado como pré-abertura. A pré-abertura equivale a um período inicial de consolidação de ofertas onde o preço é formado mas não ocorre a realização de negócios. Este período serve também para ajustar o preço das ofertas em face das primeiras intenções de investimento transmitidas para o mercado numa dada sessão de bolsa. Este período é definido pela entidade gestora de sistemas em anexo ao regulamento de negociação. No caso da Euronext Lisbon o Rule Book I.

[25] Esta funcionalidade era conseguida utilizando uma solução de processamento para um conjunto de ofertas do próprio sistema de negociação LIST, identificado como "*basket order*". A funcionalidade permitia ao *trader* responsável pela estação de trabalho configurar "*off line*" um conjunto de ofertas sobre um ou mais valores mobiliários e enviá-las para o sistema de negociação com um único comando (*click*), sendo ainda possível configurar as ofertas para que ficassem "visíveis" ao mercado somente depois de ser atingido um determinado preço.

[26] O *order collecting*, consistia, basicamente, num programa automático (*program trading*) de envio de ofertas para o sistema de negociação, que permitia aos IF nacionais reencaminharem ofertas oriundas de IF estrangeiros para o sistema de negociação (LIST). Suportado por uma plataforma *GL Trade*, esta

Em 2001, no decurso do início da negociação de um novo tipo de valor mobiliário[27], os *warrants autónomos*[28] o número de ofertas inseridas no sistema de negociação triplicou[29], duplicando subsequentemente em 2002[30]. Com o mercado de *warrants* autónomos foi introduzida a figura do *Market Maker*, que passou a promover a liquidez deste segmento assegurando, até um limite pré-estabelecido, o papel de contraparte nas operações. Para todos os *warrants* autónomos emitidos em Portugal existe um *Market Maker* registado na Euronext[31].

A introdução no contexto Euronext (2003), a adopção das diferentes categorias de Membros[32], a fusão da Euronext com a Nyse em Abril de 2007 e o arranque da DMIF em Novembro de 2007, a alteração do número máximo de casas decimais (2 para 3) nos principais valores negociados em contínuo em 2008[33] e a alteração do sistema de negociação em 2009 (NSC para UTP), transformaram o contexto de negociação do mercado de valores nacional.

Em 2008 o número médio diário de ofertas dirigidas para o mercado português ascendeu a números da ordem dos 3 milhões de ofertas, por sessão de bolsa.

Em Dezembro de 2009 o número de IF autorizados a intervir no mercado Português era de 91, sendo que apenas 18 eram nacionais[34]. Em Setembro de 2011 encontravam-se autorizados a actuar no mercado Euronext Lisbon 102 IF[35], sendo 16 nacionais.

---

solução foi inovadora para o mercado de acções em Portugal, embora já existissem redes semelhantes para o mercado monetário e obrigacionista, através das plataformas da *Reuters* e da *Bloomberg*.

[27] Os valores mobiliários admitidos à negociação eram: acções, obrigações (privadas e dívida pública), títulos de participação e pontualmente direitos de conteúdo patrimonial.

[28] Resultante da publicação do decreto-lei nº 172/99, 20 Maio e do Regulamento da CMVM nº 6/2001, Novembro.

[29] Em 2001 o número médio diário de ofertas inseridas no sistema de negociação LIST foi de 58.000.

[30] Em Novembro de 2002 o número médio diário de ofertas inseridas no LIST foi de 190.000.

[31] Número 1, artigo 10 do Regulamento 5/2004 da CMVM.

[32] A introdução de categorias diferentes de intervenientes no mercado entre Membros Negociadores, Compensadores e Liquidadores abriu aos Intermediários Financeiros não nacionais a participação directa no mercado de valores português, posteriormente identificado como *Eurolyst by Euronext*.

[33] Em 28 de Janeiro de 2008 a Euronext adoptou para o mercado Euronext Lisbon e para um conjunto de 10 valores mobiliários admitidos à negociação em Mercado Regulamentado e constituintes do indice PSI20, o tick-size de €0,005. Ver http://www.euronext.com/fic/000/030/050/300505.pdf .

[34] Em 2001 o mercado nacional contava com a intervenção de 26 Membros Negociadores, incluindo a Société Generale e o Citibank com actuações exclusivas no segmento warrants autónomos.

[35] O universo de IF autorizados a intervir no mercado português (*Euronext Lisbon*) em Outubro de 2011 era de 102. Recuperado em 08 de Outubro de 2011, de http://www.euronext.com/forourclient/mbs/market/list-1662-EN.html.

Da mesma forma o custo unitário de transacção passou de €1,35 em 2002 para 1,20€ em 2008. Em Junho de 2008, a Nyse Euronext lançou um preçário especial dedicado ao HFT designado "*Pack Epsilon*", onde o custo por negócio pode chegar aos 0,50€ (cinquenta cêntimos)[36]. O actual preçario estabelece um custo minimo de 0,60€ (sessenta cêntimos) por negócio com um cap 0.55*bp*, no entanto o rácio entre ordens e negócios é de 100:1[37]. Também a Clearnet, reduziu o custo do *clearing*, cobrando actualmente 0,05€ (cinco cêntimos) por negócio compensado.

1.2. A Relevância do Estudo

É neste novo contexto, potenciado pelo surgimento de máquinas mais potentes, maior velocidade de processamento, elevados volumes de negociação, preçários reduzidos, um maior leque de intermediários e investidores, valores decimais, menor volatilidade e menor *spread* e o recurso à actuação de máquinas suportadas em algoritmos matemáticos, que este estudo se realiza, visando identificar o estado-da-arte da negociação em acções, resultante da actuação dos IF na negociação do mercado nacional (Euronext Lisbon), e, principalmente, o seu grau de semelhança com o comportamento característico de um *High Frequency Trader* (HFTr)[38].

O estudo é ainda relevante por outros dois motivos. Em primeiro lugar, é o primeiro estudo que tenta identificar de forma estruturada o nível de actuação dos HFTr no mercado de valores português. Em segundo lugar, também este motivo de grande relevância, é o que utiliza pela primeira vez uma rede neuronal não-supervisionada para a detecção da presença do HFT na negociação de acções, diferentemente da prática

---

[36] O preço está relacionado com o número de ofertas inseridas no sistema de negociação, donde se produz um rácio entre o número de negócios realizados e o número de ofertas inseridas. O rácio utilizado no "*Pac Epsilon*" é de 30:1, ou seja, 30 ofertas para cada negócio realizado. O referido preçário não consta do Trading Fee Guide, donde se supõe descontinuado.

[37] *Trading Fee Guide*. Estrutura de taxas cobradas na negociação de acções e direitos. Recuperado em 02 de Novembro de 2011, de http://europeanequities.nyx.com/sites/europeanequities.nyx.com/files/nyse_euronext_cash_market_trading_fee_guide_1_october_2011_0.pdf, p.6.

[38] IF que utiliza a solução HFT.

utilizada no mercado que considera todos os negócios realizados via um *proprietary trading*[39].

1.3. Problema e Objectivos do Estudo

A aplicação do conhecimento de mercado, antes traduzida na actuação do *trader* no *floor* de negociação, há muito foi substituída pelos algoritmos electrónicos de negociação, que actualmente processam milhões de ofertas por segundo.

Conseguir identificar no universo de dados diários provenientes da negociação do mercado de acções nacional, a actuação de um HFTr, diferentemente de somatizar os negócios realizados via um *proprietary trading*, constituiu o objectivo principal deste estudo, na medida em que se entende ser crucial a correcta interpretação dos seus efeitos, previamente à adopção de medidas regulamentares que o suportem ou, em contrapartida, o inibam. Preocupação, aliás, patente nos trabalhos desenvolvidos no âmbito da ESMA e da IOSCO.

Neste contexto o estudo produz uma análise ao actual padrão de actuação dos IF na negociação de acções no mercado português, traduzido pelo envio de ofertas para o sistema de negociação da Euronext (Euronext Lisbon). Com recurso a uma rede neuronal não-supervisionada (*Kohonen Self-Organizing map*), é realizada uma segmentação dos IF, procurando identificar se a sua forma de actuação na negociação, lhes confere ou não, um grau de semelhança suficiente que permita equipará-los a um HFTr .

Nesta fase a questão a responder será: **O mercado regulamentado de acções português é utilizado pelos IF numa lógica de negocição do tipo HFT**?

[39] Entenda-se por *proprietary tranding* o IF que actua na negociação exclusivamente em nome próprio.

1.4. Estrutura do Trabalho de Investigação

Neste capítulo é apresentado o tema do estudo, algumas preocupações e os desenvolvimentos recentes sobre o mesmo. O mercado português é revisto no âmbito do desenvolvimento tecnológico das plataformas de negociação. Por fim, apresenta-se uma breve descrição da relevância e dos objectivos prosseguidos com a realização do estudo.

O segundo capítulo inclui o enquadramento teórico do tema. Na primeira parte é apresentada uma definição do HFT e alguns dos conceitos relacionados com a sua compreensão. É apresentada uma descrição das abordagens e metodologias utilizadas na definição e na identificação do HFT, bem como alguns dos objectivos subjacentes à sua utilização. Na segunda parte é introduzido o conceito de Redes Neuronais Artificiais, com a descrição de alguns casos de sucesso na utilização do SOM, as metodologias utilizadas para segmentação e as suas características, ambas presentes na literatura científica.

No terceiro capítulo descreve-se a metodologia seleccionada para a implementação da análise do fenómeno HFT no mercado accionista português, bem como as ferramentas informáticas utilizadas para a concretização do estudo.

O quarto capítulo é dedicado à análise dos dados, sendo apresentado o modelo de rede SOM escolhido para realizar os exercícios de treino e de validação das bases de dados.

No quinto capítulo discutem-se os resultados obtidos, sendo apresentado algumas sugestões para a investigação futura.

O sexto capítulo identifica a bibliografia consultada, sendo o sétimo capítulo reservado à descrição dos vários conceitos necessários à compreensão do HFT.

2. ENQUADRAMENTO TEÓRICO

Neste capítulo é efectuado o enquadramento teórico do tema deste estudo: o *High Frequency Trading* (HFT).

Numa primeira parte é apresentada a definição de HFT, bem como alguns dos principais conceitos necessários à sua compreensão. Suportado na revisão da bibliografia apresentamos as abordagens e metodologias utilizadas na definição e na identificação do HFT.

A segunda parte introduz o conceito de RNA, com a descrição de alguns casos de sucesso na utilização de redes neuronais não supervisiondas e ainda as metodologias utilizadas para a segmentação dos dados, ambas presentes na literatura científica. Face à característica dos dados (secundários) e ao objectivo do estudo - encontrar num universo elevado de dados semelhança de actuação tipo HFTr – optou-se pela robustez da rede neuronal *Kohonen Self-Organizing map* (SOM) para a segmentação.

2.1. High Frequency Trading

O desenvolvimento tecnológico verificado ao longo dos anos levou a maioria das estruturas de negociação mundiais a adoptarem sistemas de negociação electrónicos (ET). Na sua concepção, suportada em potentes computadores, a lógica de negócio é traduzida e incorporada num conjunto significativo de programas electrónicos (PT) que recorrem a parâmetros prévios e definidos por via regulamentarmente[40].

No contexto europeu, o fomento à concorrência entre as estruturas de negociação, implementado com a DMIF em finais de 2007, intensificou o recurso à utilização de tecnologia avançada na negociação, o que permitiu melhorar a qualidade, quer das próprias estruturas de negociação, quer dos meios de acesso (conexão) às mesmas, ambos um incentivo à adopção de estratégias de negociação suportadas em algoritmos electrónicos (AT).

[40] Regulamentação emitida pela entidade de supervisão e pelas sociedades gestoras que identificam a lógica e as directrizes do mercado e da negociação em bolsa. Em Portugal, a CMVM e a Euronext Lisbon, respectivamente.

Subproduto deste novo ambiente negociação, altamente tecnológico, o HFT está a mudar o paradigma da negociação nos mercados de instrumentos financeiros[41], onde antes a mera expectativa dos investidores e dos agentes de mercado (IF), normalmente resultante da análise fundamental do comportamento das empresas e/ou da própria economia, procurava identificar a direcção (subida ou decida das cotações), o nível de liquidez e a profundidade do mercado (*spread*). No HFT, tal resulta do mero processamento de algoritmos matemáticos em milésimos de segundo onde a principal condicionante é identificar as ligeiras "ineficiências" do mercado para obter, numa única sessão de bolsa, ínfimas, mas recorrentes mais-valias.

Associado à caracterização da solução HFT encontramos ainda um conjunto distinto de conceitos e terminologias, que importa perceber, para melhor compreender a envolvente desta investigação, nomeadamente, o ambiente propício à sua utilização, as principais estratégias de negociação que recorrem à solução e o conceito de latência, quer dos sistemas de negociação quer dos acessos (conexão) às estruturas de negociação que os suportam. A definição cuidada de cada um destes conceitos é remetida para anexo, sendo referenciada, sempre que necessário, na literatura seleccionada para melhor compreender a solução HFT.

2.1.1 Definição e principais características.

2.1.1.1. O que é high frequency trading?

O HFT é um método que utiliza tecnologia avançada (*software* inteligente) para implementar uma determinada estratégia de negociação. No entanto, o HFT não pode ser visto de forma separada como uma estratégia de negociação [42].

---

[41] O HFT é já uma realidade nos mercados de intrumentos financeiros, uma vez que ocorre não só nos mercados regulamentados de valores mobiliários mas também no mercado de derivados.

[42] "A specific type of automated or algorithmic trading is known as high frequency trading (HFT). HFT is typically not a strategy in itself but the use of very sophisticated technology to implement traditional trading strategies." (Definição apresentada pela Comissão Europeia em documento de base a consulta pública de revisão da DMIF. Recuperado em 04 de Abril de 2011, de http://ec.europa.eu/internal_market/consultations/docs/2010/mifid/consultation_paper_en.pdf, p.14).

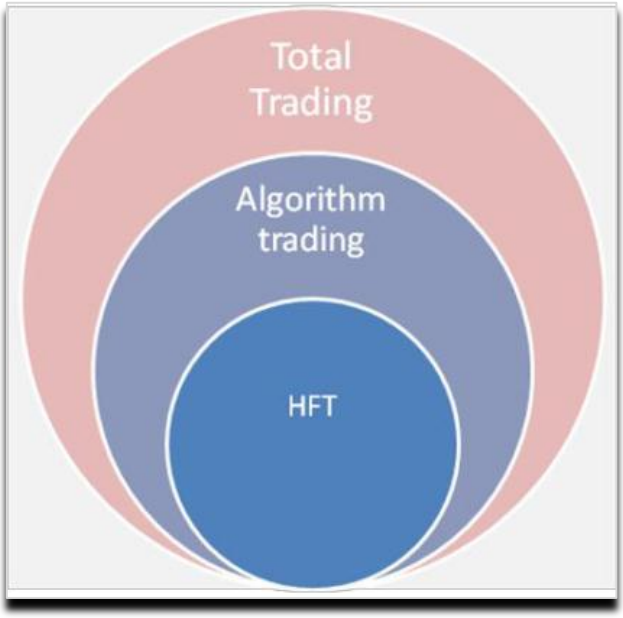


Figura 2.1. HFT como um sub-produto do algorithm trading.
Fonte: AFM, 2010, p. 11.

2.1.1.2. Principais características

Os sistemas electrónicos de negociação (ET) assemelham-se aos algoritmos de negociação (AT) quando há um computador que actua como o principal decisor na realização de uma operação, suportada em "sensores" ou "detectores" das condições do mercado[43]. No entanto, os ET e os PT são frequentemente de alta velocidade, mas não necessariamente de alta frequência. Também os AT tipicamente não envolvem volumes repetitivos e realizados em alta frequência. Em vez disso, o AT é utilizado para fomentar a liquidez fragmentada, minimizar o impacto no mercado, igualar ou melhorar um *benchmark* e melhorar a execução dos negócios.

Embora também seja visto como um subconjunto de um ET, o HFT não é sinónimo de programa de negociação (PT), ou mesmo de um algoritmo de negociação (AT). O HFT é uma forma de automatizar a realização de operações em mercado

[43] Por exemplo quando o sistema executa vários negócios resultantes de uma única oferta introduzida a preços de mercado ou ainda quando o sistema interrompe a negociação tendo em conta uma oscilação de preço acima do limite previamente definido para a sessão de bolsa ou para o instrumento financeiro.

suportado em algoritmos matemáticos[44] capacitados para decidirem, sem necessidade de intervenção humana, *quando, como* e *onde* negociar determinado instrumento financeiro[45].

As máquinas conectadas aos sistemas de negociação (ET) recebem e processam constantemente as informações do mercado[46] e decidem numa velocidade extremamente elevada, o que comprar e vender[47]. A base de todo o processo de negociação são algoritmos matemáticos, desenhados para identificar momentos específicos dos mercados, actuando sempre que se verifica uma mínima oportunidade de ganho. No entanto, o sucesso do HFT está dependente da velocidade de execução e de uma latência ultra baixa[48], esta geralmente associada à realização de um elevado número de negócios num curto intervalo de tempo.

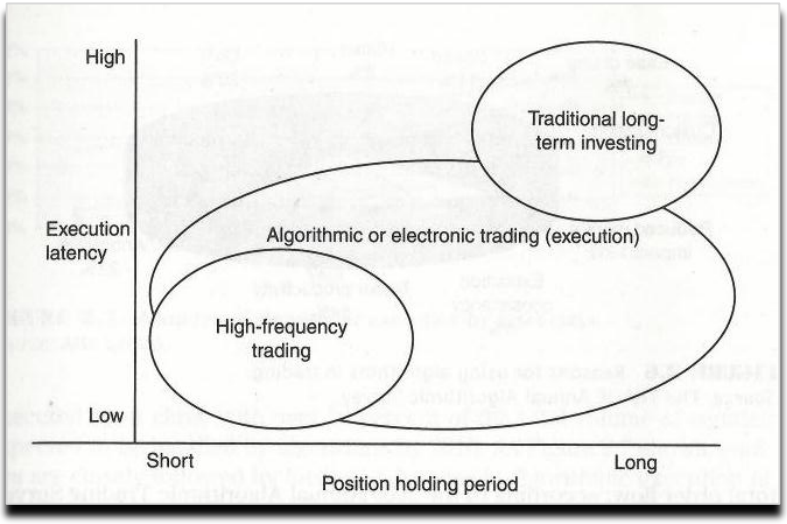


Figura 2.2. HFT versus investimento (longo prazo).

Fonte: Aldridge, 2010, p. 17

[44] Um algoritmo pode ser descrito como um conjunto de regras que, tendo em conta um certo momento da negociação (condições de mercado), executa uma série de operações.

[45] No âmbito deste estudo, valores mobiliários acções.

[46] Normalmente, os melhores preços de compra e venda (*best bid; best ask*). Conforme a estratégia a implementar esta informação poderá envolver outros mercados (eg. Mercado de derivados) ou outras plataformas de negociação (eg se a estratégia é realizar arbitragem com os mesmos instrumentos financeiros admitidos a negociação em diferentes plataformas).

[47] As estratégias de negociação utilizadas no HFT caracterizam-se por um elevado número de negócios e a realização de ganhos por negócio é em média muito baixo. O objectivo é realizar pequenos ganhos muitas vezes na mesma sessão de bolsa.

[48] Ver conceito de *Latência* apresentado no capítulo 7.

Embora a maioria dos sistemas electrónicos de negociação (ET) apresentem uma baixa *latência* na conectividade às suas bases de dados da negociação, apenas o HFT depende totalmente destas conexões com latência ultra baixa e numa base regular, uma vez que as *estratégias de investimento* e *estratégias de execução* são sinónimos neste caso. Como referido, o HFT é um método para implementar estratégias de investimento com base em algoritmos. Não existindo isoladamente, o HFT pode ser considerado apenas como mais uma componente da estrutura do mercado[49].

Outras duas situações fundamentais contribuem para a caracterização do conceito:

i) No HFT as posições abertas durante uma sessão de bolsa são maioritariamente encerradas antes do fim do dia de negociação. O objectivo principal é a realização de vários negócios, normalmente de curtíssima duração (milesegundos), com quantidades de soma nula[50]. O volume e o período de tempo em que as posições são mantidas abertas são determinados pelo algoritmo de negociação, sendo que estes podem variar durante a sessão de bolsa;

ii) muitas das ofertas enviadas para o sistema de negociação são canceladas imediatamente após o seu registo, sendo continuamente actualizadas conforme se alteram as condições do mercado. A existência de situações de *explosão* de grandes quantidades de ofertas é outra das principais características do HFT. Estas explosões, frequentemente alternam-se com períodos de relativa calma em que praticamente nenhuma negociação ocorre, em antecipação de uma nova oportunidade de negociação. O rácio obtido entre o número de ofertas introduzidas no sistema de negociação e o número de negócios (*order-to-book ratio*) realizados assume-se bastante elevado.

---

[49] Embora importante no reconhecimento do fenómeno HFT, não incluiremos no contexto deste estudo a definição de todos os "agentes" que compõem a estrutura do mercado nacional. No entanto, referimos apenas que desta constam outras formas organizadas de negociação, como as estruturas de compensação e liquidação, os mecanismos de difusão de informação de mercado, entre outros.

[50] No final da sessão de bolsa o saldo resultante entre as posições de compra e venda é igual a zero (0). Apurando-se o saldo financeiro das várias execuções.

2.1.2. O fenómento HFT.

"O *high frequency trading* é um fenómeno recente" (Brogaard, 2010), que tem gerado elevados proveitos e que está a mudar a forma de actuação dos IF na negociação em bolsa.

Como qualquer novo fenómeno, tem produzido, simultaneamente, debates, críticas e diferentes opiniões no mundo financeiro, quer nos EUA, onde primeiro se fez sentir, como no contexto europeu, agora sob revisão[51]. No entanto, como refere Brogaard (2010) "a análise académica do seu impacto nos mercados financeiros é ainda limitada", incipiente e até ao momento vocacionada a apresentar os diferentes modelos estatísticos e os vários métodos para a sua aplicação, estes que permitem implementar estratégias de investimento antes limitadas à capacidade de processamento das máquinas.

Neste sentido, e dado que o objectivo primeiro deste trabalho é produzir um modelo de identificação do HFT, neste subcapítulo efectua-se uma revisão da literatura produzida e que permita melhor caracterizar o ambiente e os incentivos associados à utilização deste conceito.

Numa primeira abordagem a este fenómeno podemos referir que "a ascensão do HFT é um resultado de duas importantes mudanças, que têm aumentado a capacidade e a conveniência a uma negociação rápida e frequente" (Brogaard, 2010). A primeira resulta da adopção do conceito de decimalização[52], que permitiu uma maior flexibilidade na gestão do *tick size*[53], tendo em conta os diferentes preços unitários das empresas (acção), bem como na difusão das intenções de compra e venda de instrumentos financeiros nas diferentes estruturas de negociação (ET), o que reflecte uma maior variação de preço a cada minuto/segundo. Como, e bem, refere Brogaard

---

[51] A Comissão Europeia e a ESMA no contexto da revisão da DMIF têm recolhido junto do mercado opiniões no sentido de clarificar o conceito, analisar as suas envolventes e seus riscos;

[52] Em 2001, no mercado americano, o valor dos instrumentos financeiros passa a estar representado pela unidade e um número de casas decimais. Algumas estruturas de negociação, como a Nyse Euronext permitem cotar valores até 3 casas decimais.

[53] Nos mercados financeiros, é o menor incremento (escala) que o preço de um instrumento financeiro (eg. acções, contratos derivados) admitido à negociação pode sofrer (p.e. € 0,001). Este incremento é normalmente definido conforme o preço unitário do instrumento financeiro, sendo sempre maior quanto maior for o preço unitário. Na revisão em curso da DMIF este é um conceito que se pretende harmonizar, uma vez que cada estrutura de negociação adopta a sua escala.

(2010) "menores incrementos no preço permitem a mudança de posição durante curtos horizontes de tempo e com um menor potencial de risco face a uma diferente variação do mercado". A segunda, e não menos importante, foi o avanço tecnológico verificado na capacidade de processamento e reacção à informação difundida no mercado, resultado do intenso fluxo de comunicação, obtido de e para as ET nos diferentes mercados. "A partir dessas mudanças, os HFTs evoluíram" (Brogaard, 2010). No seu estudo sobre a actuação dos HFTs no papel de novos *Market Makers* no mercado Belga, Menkveld (2010) refere que as actuais máquinas, à semelhança dos *brokers*, "têm habilidade para processar grandes quantidades de informação (de mercado), no entanto fazem-no quase instantaneamente"[54].

Este estado de eficiência, no processamento e execução de uma oferta, tem início com a intensificação no uso dos algoritmos electrónicos, superando a sua então passiva utilização de transmissão de blocos de ofertas para as estruturas de negociação[55]. Numa análise à negociação de um conjunto acções admitidas à negociação na Nyse, Hendershott, e*t al.* (2010) concluíram que "os custos decrescentes da tecnologia levaram à adopção generalizada dos AT em toda a indústria financeira", tendo como resultado uma revolução na forma como os agentes de mercado (IF) passaram a actuar em bolsa. "Muitos IF realizam agora as suas operações via algoritmos electrónicos" contribuindo, segundo o estudo, para o aumento da liquidez do mercado. "Nos cinco anos seguintes à decimalização, a utilização do AT aumentou, e os mercados tornaram-se mais líquidos" (Hendershott, et al., 2010)[56].

Com a intensificação do uso de algoritmos electrónicos aumentou a procura de uma cada vez menor latência que permitisse implementar estratégias de "investimento" até então pouco exequíveis. Hasbrouck e Saar (2011), utilizando uma amostra do livro de ofertas da Nasdaq, verificaram que "muitos IF respondem a pequenas variações nos preços dos activos, alterando o limite do livro de ofertas central (*best bid/ best ask*) em média em cerca de 2 a 3 ms (milissegundos) ". Os autores definem como "actividade de

---

[54] Vide Menkveld (2010), "High Frequency Trading and The New-Market Makers", p. 3.

[55] Como referido anteriormente os algoritmos electrónicos serviam de base aos chamados "*order collecting*", que cumpriam uma função de transmissão de um conjunto de ofertas para as estruturas de negociação, previamente programadas, no sentido de dar cumprimento às intenções de investimento dos clientes dos IF (eg. bancos, grandes corretores). Estas ofertas eram normalmente transmitidas para o sistema de negociação antes do início da negociação e com diferentes prazos de validade.

[56] Vide Hendershott et.al. (2010) "Does Algorithmic Trading Improve Liquidity?", pp. 30, 31.

baixa latência (*low latency*) as estratégias que respondem a eventos do mercado no milissegundo", identificando-as como a "imagem de marca" dos IF que actuam por conta própria (*proprietary trading*) face aos demais *players* (IF) do mercado. A métrica utilizada pelos autores resulta da análise de estratégias dos IF associadas à submissão e respectivo (e imediato) cancelamento da oferta e às execuções susceptíveis de constituírem parte de uma qualquer estratégia (eg. market making, arbitrage). As conclusões do referido estudo apontam para um incremento da qualidade do mercado, como sejam a redução da volatilidade no curto prazo e um aumento na profundidade do livro de ofertas, com o aumento da actividade de baixa latência (*low latency*).

Em consonância com a exigência de optimização do mercado (menor latência) e do aumento da concorrência resultante do surgimento de novas estruturas de negociação (MTF), introduzidas com a DMIF, as sociedades gestoras de mercado regulamentado (bolsas), também contribuíram para a redução da latência das suas estruturas de negociação (ET). Numa análise à estrutura de negociação do mercado regulamentado alemão (Deutsche Boerse), Riordan e Storkenmaier (2008), verificaram que o upgrade realizado ao sistema Xetra[57,] "contribuiu dramaticamente para o aumento da liquidez naquele mercado, reduzindo a latência de 50ms para 10ms". Hendershott e Moulton (2009), examinaram as alterações no mercado regulamentado norte-americano, especificamente o livro de ofertas da Nyse após o *upgrade* do sistema de negociação realizado em finais de 2006. Hendershott e Moulton (2009), identificaram significativas mudanças na qualidade do mercado tendo em consideração o nível da tecnologia, verificando que "o tempo de execução das ordens de mercado reduziu de 10 segundos para menos de um segundo, tendo-se verificado também uma redução no intervalo de preços entre as melhores ofertas de compra e venda (*bid/ask spread*)", ou seja "uma mais eficiente formação de preços".

Como complemento à caracterização do ambiente de actuação do HFTs, e porque deste depende o objectivo do estudo proposto, importa melhor identificar como a literatura disponível se refere às diferentes tipologias de agentes (IF) no mercado de

[57] A versão 8.0 do sistema de negociação Xetra arrancou em 23 de Abril de 2007. A Nyse Euronext, de que faz parte a Euronext Lisbon, estrutura de negociação utilizada para este estudo realizou o upgrade do seu sistema de negociação NSC para o UTP em Fevereiro de 2009. O actual tempo de latência do UTP é de 10ms.

acções face à sua actuação junto dos sistemas electrónicos de negociação (ET). Face ao cenário tecnológico referido, é fácil perceber que "muita da actividade verificada nos mercados accionistas (eg. EUA) é comummente atribuída à negociação algorítmica". No entanto, como referem Hasbrouck e Saar (2011), "nem todos os algoritmos têm a mesma finalidade e, portanto, os padrões que eles induzem aos dados e o impacto que têm sobre a qualidade do mercado dependem de seus objectivos específicos". De um modo geral, até porque é assim que as estruturas de negociação (ET) os permitem classificar, a actividade dita "algorítmica" é identificada ou como de "agência" (ou para clientes) ou como de "proprietários" (*proprietary trading* - os que actuam em nome próprio). Os algoritmos de agência (AA), são utilizados por clientes dos IF membros do mercado, normalmente com o objectivo de minimizar o custo de execução de negócios no processo de implementação de mudanças nas suas carteiras de investimento[58]. Os algoritmos proprietários (AP) são utilizados por "criadores electrónicos do mercado (*Market Maker*), *hedge funds*, mesas de negociação de grandes IF, e empresas independentes de arbitragem estatística, e destinam-se a lucrar com o ambiente de negociação própria " (Hasbrouck e Saar, 2011). Esta estratégia diverge substancialmente da tradicional negociação realizada com base em análise fundamental.

Importa notar que a investigação relativa às estratégias de utilização da solução HFT não é referida neste subcapítulo, uma vez que o objectivo principal do trabalho consiste em conseguir identificar o grau de utilização da tecnologia no mercado accionista português, e não a melhor forma de utilizá-la para obtenção de mais-valias financeiras, num contexto específico de mercado.

2.2. Redes Neuronais Artificiais (RNA)

O tema abordado neste estudo permite antever que a inovação tecnológica e o custo cada vez menor na negociação electrónica de instrumentos financeiros estão a revolucionar a negociação em bolsa. Neste contexto e à semelhança das empresas

[58] Ver os conceitos de DMA e SA, que poderão ser utilizados, por exemplo por fundos de investimentos, fundos de pensões na gestão das suas carteiras de investimentos. A estratégia normalmente adoptada neste tipo de algoritmos electrónicos é a de adquirir ou vender a quantidade pretendida com a menor variação possível no preço e causando o menor impacto no mercado, e não execução de várias operações num curto intervalo de tempo.

admitidas à negociação nos seus mercados ou sistemas, que procuram cada vez mais diferenciar os seus produtos e serviços, obtendo assim vantagens competitivas face à concorrência[59], também as sociedades gestoras de sistemas de negociação multilaterais (bolsas) encontram-se a braços com esta diferença[60], proporcionando aos diferentes agentes do mercado, soluções tecnologicamente inovadoras e competitivas em termos de custos.

Assim, à semelhança do que sucede com as organizações em geral, a obtenção de informação relevante e atempada necessária ao controlo e à supervisão do cumprimento das regras de mercado, num ambiente com um exponencial volume de dados electrónicos, é hoje somente possível com recurso aos sistemas de informação.

Na literatura académica, a identificação do HFT num ambiente com um crescente volume de dados não foi ainda associada a uma metodologia específica, limitando-se apenas a somatizar a intervenção dos *proprietary trading*. Face às características dos dados secundários[61], entendemos abordar o tema utilizando técnicas de modelação descritiva.

A modelação descritiva "concentra-se na descoberta de padrões descritivos dos dados que sejam passíveis de serem compreendidos e interpretados", e "consiste em separar um conjunto de dados expresso num espaço p-dimensional em grupos, tanto quanto possível, homogéneos" (Bação, 2007 – cap.I, p.22).

Do ponto de vista metodológico existem várias técnicas estatísticas de redução da dimensionalidade e interpretação dos dados. No entanto, "os métodos convencionais em estatística são capazes de revelar as regularidades, tendências e estruturas em dados brutos mas, poucos métodos permitem visualizar directamente as relações entre os

---

[59] Esta é uma afirmação que resulta da compreensão da lógica do mercado de instrumentos financeiros. É facto que, identificar as inovações introduzidas por cada uma das 20 empresas admitidas à negociação na Euronext Lisbon, constituintes do principal indicador de mercado português o PSI20, obrigar-nos-ia a uma análise de pormenor aos "Relatórios e Contas" das respectivas sociedades, algo que não faz parte do escopo desta investigação. No entanto, não é difícil de assumir que empresas como a EDP, a Portugal Telecom ou qualquer dos grupos bancários (BES, BCP, BPI e SANTANDER), não estejam activamente a procura de novas e melhores soluções para os seus clientes, o que consequentemente resulta numa maior eficiência do sector e do próprio mercado de capitais.

[60] Atente-se nas várias alterações ocorridas nas estruturas de negociação nos principais sistemas multilaterais, quer na América quer na Europa.

[61] Os dados seleccionados para a identificação do problema (HFT), não foram recolhidos com o propósito específico de responder a uma determinada questão mas, resultam antes do quotidiano de negociação do mercado de valores mobiliários português – Euronext Lisbon.

elementos em grandes e complexos conjuntos de dados" (Deboeck & Kohonen, 1998, p.159).

No universo de técnicas de Data Mining encontramos as habitualmente designadas Redes Neuronais Artificiais (RNA), normalmente sub-divididas em dois tipos principais e segundo os algoritmos de aprendizagem (treino): i) Não-supervisionadas e ii) Supervisionadas. "O algoritmo de aprendizagem (treino) designa-se supervisionado quando, para a fase de treino do modelo, o *output* é conhecido previamente". "No treino das RNA não-supervisionadas não existe variável de *output*" (Bação, 2007, cap. VII, p.7).

"Uma RNA pode ser encarada como uma representação computacional do que pensamos ser o cérebro humano e o seu funcionamento, sendo o principal desafio simular, através de representações computacionais, o seu processo de aprendizagem" (Bação, 2007, cap. VII, p. 1).

À semelhança do processo de difusão de informação de um neurónio biológico, na rede neuronal artificial, a lógica associada à sua interpretação "consiste essencialmente num conjunto de unidades de processamento simples (neurónios) que comunicam entre si enviando sinais através de um número elevado de conexões" (Horta e Mendes, 2007, p.5). "A anatomia e o funcionamento do neurónio artificial são muito semelhantes à de um neurónio natural" (Almeida, 2011, p. 19), no entanto "o cérebro humano não é mais do que uma inspiração, muitas vezes remota, para as redes neuronais" (Bação, ibid., p.1).

As redes neuronais artificias (RNA), ao contrário das ferramentas estatísticas tradicionais (p.e. regressão), "abordam o problema sem qualquer pressuposto ou limitação de complexidade" (Bação, 2007, cap.VII, p.4). Esta propriedade torna-as particulamente úteis quando o conhecimento das relações das váriaveis a modelar é incipiente.

As RNA apresentam características importantes, tais como: "o aprendizado e a generalização a partir de um conjunto de dados, e a aproximação de funções contínuas multi-variáveis lineares e não-lineares, que fazem delas uma ferramenta atractiva na tarefa de modelagem e previsão de séries não estacionárias" (Neto, 2008, p.5).

Face ao objectivo deste estudo[62], considera-se que as RNA permitirão melhor conhecer o nível de ofertas resultantes da actuação dos HFTs na negociação de acções no mercado português, constituindo assim uma importante informação quer para as sociedades gestoras de sistema[63] quer para os organismos de supervisão do mercado, na verificação e/ou introdução de regras que garantam o normal equilíbrio da negociação em bolsa e a análise e verificação de práticas de manipulação de mercado.

Neste sentido, e considerando o relevante desempenho conseguido em diferentes sectores da economia, optamos por utilizar uma RNA não-supervisionada do tipo *Kohonen Self-Organizing map* (SOM) para identificar os IF com grau de semelhança suficiente a equipará-los a um HFTr (clustering).

2.2.1. Self-Organizing map (SOM).

"Apesar do termo *Self-organizing map* ser aplicado a um grande número de abordagens" (Bação, 2007, cap.V, p.1), adoptou-se a referência ao termo como sinónimo de *Kohonen Self-organizing map*. "*Konhonen Self-Organizing map* (SOM) é uma rede neuronal não-supervisionada que mapeia os n-dimensional dados de *input* para um mapa de *output* (espaço de *output*) de menor dimensão, preservando porém as relações topológicas originais" (Kiang at al., 2006, p. 37). "A propriedade de preservação da topologia significa um grupo vetores de entrada semelhante sobre os neurónios: pontos que estão próximos uns dos outros no espaço de entrada são mapeados para as unidades próximas no mapa de *output* do SOM" (Deboeck & Kohonen, 1998, p.xxxiv).

A sua característica "não linear", permite à rede ser "treinada de forma a aprender ou a encontrar relações entre os dados (vectores) de *input* e *output* ou a apresentar os dados de forma a se perceber neles padrões desconhecidos" (Almeida, 2011, p.21), sendo o "raio de vizinhança o parâmetro que garante a preservação da estrutura topológica" (Bação, 2007, cap.V, p. 4). O "treino do SOM é governado por um

[62] Encontrar num universo de dados secundários (ofertas inseridas no sistema de negociação) um conjunto específico que se assemelhe a solução HFT (segmentar) e posteriormente produzir um modelo que permita identificar a quota de mercado das operações resultantes da referida solução (classificar).

[63] Por exemplo na identificação de preçários especiais.

parâmetro, normalmente designado taxa de aprendizagem, que em cada momento define o ajustamento que é feito" na rede (Bação, 2007, cap.V p.5; Deboeck & Kohonen, 1998, p. 164).

Por outras palavras, o SOM "é uma rede neuronal com capacidade para organizar grandes quantidades de dados de acordo com as semelhanças que apresentam" (Bação, 2007, cap.V, p. 2), ou seja, uma excelente "ferramenta" de agrupamento e/ou de visualização de dados de alta dimensão.

O SOM, à semelhança de outros algoritmos de agrupamento, recorre a "heurísticas que atribuem objectos a grupos com base em medidas de distância entre o objecto e o centróide do cluster". Segundo Kiang at al. (2006, p. 38), estes algoritmos "não são estatísticos, no sentido de que eles não dependem de quaisquer suposições de distribuição". [64]

2.2.1.1. O treino do SOM

Figura 2.3. Self-organizing map

Fonte: Deboeck & Kohonen, 1998, p. xxxv

"O processo de treino do SOM é muito simples" (Bação, 2007, cap.V, p. 5). Apresentado um padrão de input à rede, "procede-se à avaliação das distâncias entre o

[64] Para um maior detalhe do processamento do algoritmo *Self-organizing map* ver Deboeck & Kohonen (1998) - *Visual Explorations in Finance with Self-Organizing Maps*, p. 160 e segs..

objecto (vector) e todas as unidadades" da camada de *output* , "sendo que a mais próxima ganha a representação do indivíduo" (Bação, ibid., p.5). "A unidade de output vencedora será aquela cujos pesos ou ponderações nas conexões de entrada são os mais próximos do padrão de *input* em termos de distância euclidiana" (Almeida, 2011, p. 23).

Definida a unidade vencedora, esta mover-se-á no sentido do objecto, ajustando a sua representação face ao mesmo. A magnitude da aproximação, entre o objecto e a unidade SOM, é definida pela "taxa de aprendizagem", que pode variar no intervalo entre 0 e 1. " Tipicamente, a taxa de aprendizagem é iniciada com valores próximos de "1", tendendo para "0" durante o processo de aprendizagem (Deboeck & Kohonen, 1998, p. 162), este que decorre até que todos os padrões de *input* sejam apresentados (processados). O ciclo de processamento dos padrões de *input* é identificado, na linguagem das redes neuronais, como época. Ao fim de cada época os parâmetros do processo, taxa de aprendizagem e "vizinhança" são actualizados. "A duração do processo de treino de um SOM é definida à partida pelo utilizador, sendo tipicamente na ordem dos milhares" (Bação, ibid., p.5).

2.2.1.2. Diversidade de soluções

Ao contrário do verificado com o HFT, a literatura académica sobre a utilização de redes neuronais não-supervisionadas na área financeira é vasta, seja na classificação de diferentes empresas num dado sector da economia, na análise dos resultados das empresas, na identificação de grupos de consumidores/clientes, na classificação de títulos de dívida, na identificação de semelhanças entre diferentes países, entre outros. Importa por isso reter que a rede SOM permite a sua aplicação a uma diversidade de soluções, onde o objectivo principal seja identificar um grupo homogéneo e de elevado grau de semelhança. A revisão realizada neste subcapítulo aborda apenas um limitado mas, suficiente, conjunto de literatura que permite identificar a abrangência desta solução.

Almeida (2011), recorre à utilização do algoritmo *Self-Organizing map* (SOM), para caracterizar as Organizações Publicas Portuguesas, utilizando "rácios que provêm de mapas de contabilidade orçamental, estes que medem a execução da despesa e da receita face ao previsto no orçamento".

Kiang at al., (2004), suportado em factores demográficos e no consumo dos utilizadores da *American Telephone and Telegraph Company* (AT & T), utilizam o SOM para definir segmentos de consumidores, segundo o seu tipo de utilização em chamadas telefónicas.

Kumara e Ravi (2006), "com base em observações realizadas entre 1968-2005, apresentaram uma revisão detalhada e um conjunto de técnicas passíveis de serem utilizadas na resolução questões relacionadas com a previsão de falências enfrentadas por bancos e empresas", sendo uma dessas técnicas as redes neuronais. Os problemas de insolvência de empresas são também analisados por Neves e Vieira (2004).

Deboeck & Kohonen (1998), apresentam na obra *Visual Explorations in Finance with Self-Organizing Maps*, uma colectânia de diferentes soluções de utilização do Self-organizing map, abrangendo áreas distintas como: 1.a. Análise de demonstrações financeiras e informações para a formulação da estratégia corporativa, 1.b. Diagnóstico visual da situação financeira das empresas, 1.c. Classificações de ratings de obrigações, 1.d. Análise da convergência económica dos países europeus suportados em indicadores macroeconómicos, 1.e. SOM como sistema de apoio à decisão[65]; 2. Análise da distribuição da taxa de juro das obrigações[66]; 3. Análise da performance dos fundos mútuos[67]; 4. Descoberta de padrões subtis de transtorno nas empresas com base na demostração de resultados[68]; 5. Análise de Instituições bancárias[69]; 6. Nível de investimento nos mercados emergentes[70]; 7. Negociação em mercado de acções[71]; 8. Preço de terrenos no mercado Finlandês[72]; 9. Na avaliação dos preços de habitações[73]; 10. Identificação de clusters de consumidores no mercado Chines[74].

---

[65] 1. Cinca (pp. 3, 23). "Let Financial Data Speak for Themselves".

66 2. Bodt at al., (pp. 24, 38). "Projection of Long-term Interest Rates with Maps".

67 3. Deboeck, Guido (pp. 39, 58). "Piking Mutual Funds with Self-organizing Maps".

68 4. Kiviluoto & Bergius, P. (pp. 59, 71). "Maps for Analyzing Failures of Small and Medium-sized Enterprises".

69 5. Shumsky & Yarovoy (pp. 72, 82). "Self-organizing Atlas of Russian Banks".

70 Deboeck (pp. 83, 105). "Investment Maps of Emerging Markets".

71 Resta (pp. 106, 116). " A Hybrid Neural Network System for Trading Financial Markets".

72 Carlson (pp. 117, 127)."Real Estate Investment Appraisal of Land Properties using SOM".

73 Tulkki (pp. 128, 140). "Real Estate Investment Appraisal of Buildings using SOM".

74 Schmitt & Deboeck (pp. 14, 156). "Differential Patterns in Consumer Purchase Preferences using Self-organizing Maps: A Case Study of China".

A revisão da literatura académica e as questões suscitadas, quer pelos reguladores, quer pelo próprio mercado, identificam as preocupações desta "revolução silenciosa", que contrasta com a negociação fervorosa dos tradicionais *floor* de negociação. O HFT encontra-se efectivamente a mudar o paradigma da negociação em mercado de instrumentos financeiros.

Neste segundo capítulo realiza-se um enquadramento do tema em estudo – o HFT. Apresenta-se uma definição do HFT, suas características e as abordagens metodológicas utilizadas, quer quanto à sua definição, quer quanto a sua mensuração. Refere-se ainda algumas das preocupações relacionadas com o tema. Posteriormente, é realizado uma breve descrição das redes neuronais e do SOM, solução adoptada para a segmentação dos IF com nível de actuação equivalentes a um HFTr no mercado português de acções.

No próximo capítulo aborda-se as questões metodológicas do estudo: uma breve caracterização da base de dados, as preocupações com o pré-processamento dos dados e a sua análise descritiva e a base de dados utilizada para o treino e a validação do SOM. Especificamos os métodos utilizados para a segmentação e finalizamos com as ferramentas utilizadas no estudo.

3. METODOLOGIA

Como referido o objectivo do estudo é compreender em que medida o HFT faz já parte, ou não, da realidade portuguesa. Neste sentido analisámos o padrão de actuação dos IF no envio de ofertas para o sistema de negociação da Euronext relativamente à negociação de acções no mercado português (Euronext Lisbon).

A análise, suportada no *output* de uma rede SOM, visa produzir uma segmentação dos IF segundo a sua forma de actuação na negociação, procurando assim identificar se as ofertas por estes transmitidas ao sistema de negociação possuem um grau de semelhança suficiente para equipará-los a uma actuação exclusiva de um HFTr.

Neste capítulo apresenta-se a metodologia utilizada na análise do fenómeno HFT no mercado de acções português, bem como as ferramentas informáticas utilizadas para o efeito.

No primeiro ponto descorremos sobre a informação constante das base de dados originais obtidas junto da Euronext Lisbon, a caracterização dos dados relativos às ofertas e aos negócios, os contrangimentos e as opções adoptadas quanto ao tratamento da informação. Por fim, realiza-se uma breve estatística descritiva dos dados de forma a melhor conhecer a relevância dos mesmos.

No segundo ponto são apresentadas as diferentes fases metodológicas de processamento dos dados, reflexo do objectivo seguido ao longo da realização desta investigação. Efectua-se seguidamente a análise descritiva das variáveis relevantes para análise do problema e a opção adoptada para a normalização dos dados.

No terceiro ponto, apresentam-se as bases de dados utilizadas no treino e na validação da rede SOM.

No quarto e último ponto identificam-se as ferramentas informáticas utilizadas na execução das diferentes tarefas inerentes à investigação.

3.1. Análise descritiva dos dados.

3.1.1.Caracterização da Base de Dados original[75].

Os dados utilizados neste estudo foram extraídos da base de dados da Euronext Lisbon, compreendendo informação relativa à negociação[76] de 12 acções pertencentes ao cabaz de títulos constituintes do índice PSI20, principal indicador do mercado de valores mobiliários português.

Face ao contexto do estudo, a selecção dos referidos instrumentos atendeu a um critério único de selecção, o número médio de ofertas[77] registadas no sistema de negociação da Euronext, por sessão de bolsa, durante o primeiro semestre de 2011. A amostra inicial, utilizada para a identificação das variáveis relevantes, é no entanto, relativa à negociação realizada durante o mês de Agosto de 2011[78].

3.1.1.1. Base de Dados de Ofertas

A base de dados de ofertas identifica as intenções de investimento transmitidas ao sistema de negociação da Euronext pelos agentes de mercado autorizados a actuarem em nome próprio (*proprietary trading*) ou pelos seus clientes. A informação recebida da Euronext Lisbon, compreende os seguintes campos.

Quadro 3.1 – *Base de dados de Ofertas - Variáveis Originais*

Variáveis	Descrição
Dat_Ses	Data da Sessão de bolsa
ISIN	Código de identificação do instrumento[79]
Cod_If	Código de identificação do Intermediário Financeiro[80]

(continua)

[75] A característica das tabelas que compõem a base de dados é apresentada anexo (ponto 7.2).
[76] Entenda-se "informação relativa à negociação" como sendo o conjunto de ofertas transmitidas ao sistema de negociação da Euronext e os negócios destas derivados.
[77] Na lógica do mercado português "oferta" identifica, o registo da intenção de um investidor, realizado pelo IF junto do sistema de negociação. A intenção do investidor, comunicada ao IF é antes identificada com a terminologia "ordem".
[78] Ao contrário do normalmente verificado no mercado de valores português o referido período de negociação foi bastante activo, reflectindo o momento económico vivido na maioria dos mercados europeus.
[79] No caso em estudo o código de identificação das acções.
[80] Código atribuído pela Euronext ao IF, utilizado para identificação no sistema de negociação.

Quadro 3.1 – *Ficheiro de Ofertas -Variáveis Originais (continuação)*

Ord_Stat	Estado da oferta inserida na sessão de bolsa: 0-New; 1-Partially filled; 2-Filled 3-Done for Day; 4-Cancelled; 5-Replaced; 8-Rejected; C-Expired; S-Cancelled by Market Operation; 0-Eliminated by Corporate Event.
Ord_Time	Hora de registo da oferta no sistema de negociação (HHMMSS999999).
Cnc_Dat	Hora de cancelamento/modificação/execução da oferta (HHMMSS999999).
Ord_Ori	Origem da oferta: 1-Client, 2-House 6-Liquidity Provider, 7-Related Party, 8-Riskless Principal.
Val_Typ	Prazo de validade da oferta: 0-Day, 1-GTC, 2-VFA, 3-IOC, 4-FOK, 6-GTD, 7-VFC; V- good for VWAP.
Ord_Typ	Tipo de oferta: 1-Market, 2-Limit, 3-Stop, 4-StopLimit, P-Pegged, K-Market to limit
Qtd_Ord	Quantidade de títulos registados para oferta.
Pre_Ofer	Preço da oferta
Ord_id	Número da oferta identificada pelo sistema de negociação.

3.1.1.2. Base de Dados de Negócios

Identifica os negócios realizados, por sessão de bolsa, resultantes do encontro das ofertas introduzidas no sistema de negociação e para um dado instrumento financeiro (acção). A informação recebida da Euronext Lisbon, compreende os seguintes campos.

Quadro 3.2 – *Base de dados de Negócios -Variáveis Originais*

Variáveis	**Descrição**
Dat_Ses	Data da Sessão de bolsa
Cod_ISIN	Código de identificação do instrumento
Num_Not	Número do negócio

(continua)

Quadro 3.2 – *Base de dados de Negócios -Variáveis Originais (continuação)*

Hor_Neg	Hora do negócio
Pre_Neg	Preço do negócio
Qtd_Neg	Quantidade do negócio
Dat_Ordc	Data da oferta de compra
Dat_Ordv	Data da oferta de venda
Num_Ordc	Número da oferta de compra
Num_Ordv	Número da oferta de venda

3.1.1.3. Constrangimentos e procedimentos iniciais realizados sobre a informação transmitida

A ocorrência do HFT é sempre verificada em pequenos intervalos de tempos numa qualquer sessão de bolsa. Neste sentido, e como facilmente se depreende, a correcta mensuração do tempo de processamento das ofertas registadas no sistema de negociação é uma informação crucial à boa prossecução do estudo.

No entanto, logo se verificou que a base de dados de ofertas não apresentava, para o campo "Cnc_Dat", qualquer informação (hora do processamento) sempre que o estado da oferta ("Ord_Stat") inserida no sistema de negociação, fosse:

- 0-*New* – oferta inserida e não executada.
- 1-*Partially filled* – oferta parcialmente executada.
- 2-*Filled* – oferta executada (satisfeita).
- 3-*Done for Day* – oferta válida somente para a sessão de bolsa.
- C-*Expired* – oferta válida até a referida sessão de bolsa.
- S-*Cancelled by Market Operation* – oferta cancelada por acção da Euronext Lisbon.

A identificação prévia deste constrangimento obrigou a realizar um procedimento adicional de alteração da informação, antes do início do processo de análise exploratória dos dados. Assim, sempre que uma oferta apresentava o estado "0"

ou "C" ou "S", atribuiu-se ao campo "*Cnc_Dat*" uma hora igual a 18:00:00[81]. A informação (hora do processamento) relativa às ofertas com estado "1", "2" ou "3" foi obtida com recurso a base de dados da negociação, através da correspondência do respectivo número de oferta.

Ultrapassado o constrangimento da falta de informação, procedeu-se a necessária codificação[82] das acções e dos intervenientes na negociação.

3.1.2. Análise exploratória de dados

O elevado volume de dados obrigou à realização de uma análise exploratória prévia, no sentido de seleccionar a informação que melhor poderia caracterizar o grau de semelhança a uma actuação de um HFTr. Como expectável, importa analisar principalmente as acções onde se verificam um elevado volume de ofertas registadas no sistema de negociação. Neste sentido, começou por se analisar a informação relativa às ofertas registadas no sistema de negociação para as 12 acções seleccionadas[83].

Tendo em consideração o universo de agentes autorizados a actuar no mercado português (Euronext Lisbon)[84], produziu-se uma breve estatística sobre o grau de relevância destes (agentes) na negociação das acções seleccionadas.

Verificou-se que (vide figura *infra*), para a maioria das acções seleccionadas existe uma significativa concentração da negociação em um reduzido conjunto de IF[85].

---

[81] Momento seguinte ao término do período de negociação e da negociação fora de horas (*pós-trading*), onde se depreende ser o momento a partir do qual se iniciam os procedimentos de actualização da informação no sistema de negociação.

[82] A codificação das acções e dos intervenientes na negociação (IF) foi condição *sinequanom* para a utilização da informação obtida junto da Euronext Lisbon.

[83] O período temporal utilizado neste primeiro exercício foi o mês de Agosto de 2011.

[84] O universo de agentes autorizados a intervir no mercado português (*Euronext Lisbon*) compreende actualmente 102 IF. Do universo de IF, 15 apresentam mais do que um código de negociação (número máximo por IF igual a 6). Recuperado em 08 de Agosto de 2011, de http://www.euronext.com/forourclient/mbs/market/list-1662-EN.html.

[85] O número total de IF presentes na negociação dos valores seleccionados foi cerca de 9,8% do total de 102 agentes autorizados a actuar na negociação da *Euronext Lisbon*.

Inst Financeiro	N IF Partic	% IF Partic. Negociação	Nº Ofertas Top10	Nº Ofertas Resto M	N Total Ofer	Ofer Exec Inst e Mls	% Ofer Mls
A	**71**	**95,95%**	**92,28%**	**7,72%**	**1.520.488**	**625.229**	**41,12%**
B	**63**	**85,14%**	**85,54%**	**14,46%**	**529.953**	**213.142**	**40,22%**
C	64	86,49%	66,05%	33,95%	175.610	37.109	21,13%
D	64	86,49%	71,32%	28,68%	265.272	73.619	27,75%
E	**58**	**78,38%**	**90,69%**	**9,31%**	**452.182**	**255.486**	**56,50%**
F	68	91,89%	74,78%	25,22%	371.413	80.165	21,58%
G	64	86,49%	79,10%	20,90%	588.500	183.705	31,22%
H	63	85,14%	78,05%	21,95%	417.718	114.081	27,31%
I	**50**	**67,57%**	**94,56%**	**5,44%**	**297.733**	**174.190**	**58,51%**
J	**51**	**68,92%**	**90,64%**	**9,36%**	**218.983**	**97.327**	**44,45%**
M	**56**	**75,68%**	**88,94%**	**11,06%**	**483.280**	**289.146**	**59,83%**

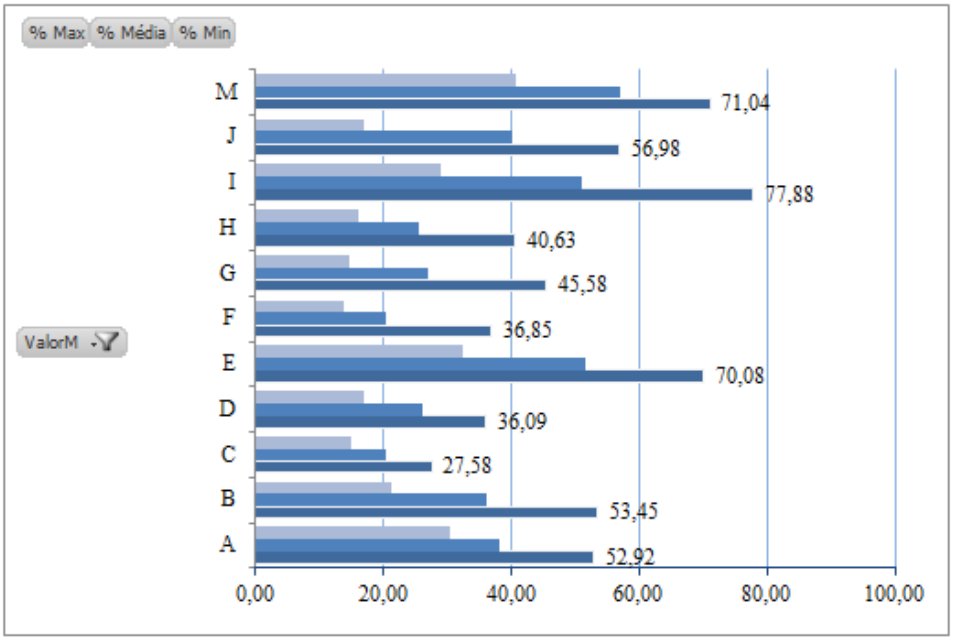
Figura 3.1 – Nível de participação na negociação - Agosto de 2011.

Fonte: informação produzida com base no ficheiro de ofertas.

Nota: Top10IF - igual a 10 agentes (IF).

Uma segunda medida produzida, também esta de elevada significância para o estudo, foi verificar o nível de relevância das ofertas processadas no milésimo de segundo, face ao total de ofertas inseridas numa dada sessão de bolsa e para um valor mobiliário[86].

[86] Esta medida é obtida pela confrontação (diferença) entre a informação do campo "*Cnc_Dat*" com o campo "*Ord_Time*".



% Min, Média e Máxima do nível de participação de um IF nas ofertas processadas no milésimo de segundo face ao total de ofertas inseridas na sessão de bolsa.
Período de análise : Agosto, 2011.

Figura 3.2 – Nível de participação na negociação - Agosto de 2011.
Fonte: informação produzida com base no ficheiro de ofertas.

Verificou-se que para as acções onde o nível de participação dos 10 principais intervenientes na negociação (TOP10IF) era superior a 85%, a percentagem máxima verificada para as ofertas processadas no milésimo de segundo era superior a 52,9%.

Neste sentido, optou-se por seleccionar para a análise de pormenor apenas seis das 12 acções, tendo estendido o período de análise para 4 meses de negociação. (cerca de 86 sessões de bolsa, realizadas entre Junho e Setembro de 2011).

O passo seguinte da análise exploratória dos dados, revelou um segundo problema. A principal variável de interesse, o estado final da oferta no sistema de negociação (campo "*Ord_Stat*") é uma variável categórica[87]. Uma vez que as demais variáveis de interesse, também eram categóricas, houve a necessidade de criar novas variáveis, estas intervalares, de forma a reflectir a informação necessária ao treino da rede SOM.

[87] A referida variável pode identificar até 10 estados diferentes para as ofertas registadas no sistema de negociação.

Importa neste ponto referir, e como poder-se-á verificar a seguir, que a actuação como HFTr de um qualquer IF na negociação poderá também ocorrer em um pequeno intervalo de tempo e durante uma única sessão de bolsa. Não constituindo este, normalmente o padrão de actuação do HFTr, a ocorrer, poderá não ser detectado pelo estudo, dado que optou-se por direccionar o trabalho às acções com maior volume de dados.

3.2.Pré-processamento de dados.

O "objectivo fundamental da fase de pré-processamento dos dados consiste em facilitar e simplificar o problema a tratar, sem excluir ou danificar informação importante para a modelação e para o entendimento do problema" (Bação, 2007, cap. III, p. 14).

Como refere Bação (2007, cap.III, p.22), é na fase de pré-processamento dos dados que o "conhecimento do analista sobre o domínio da aplicação" exerce "um aspecto fundamental para o sucesso de todo o exercício. A sensibilidade à importância das diferentes variáveis e o modelo lógico que suporta o entendimento que existe sobre o fenómeno de interesse dificilmente pode ser completamente substituído por meios automáticos".

> Uma forma de potenciar o conhecimento apriorístico do analista, consiste na construção de indicadores/índices significativos para o fenómeno a estudar. A construção de indicadores constitui uma forma eficiente de proporcionar conhecimento facilmente acessível à ferramenta de modelação e pode também permitir a redução do espaço de *input*. Obviamente, o tipo específico de indicadores a utilizar é largamente dependente do domínio de aplicação. (Bação,cap.III, ponto 3.1.1 Conhecimento Específico do Domínio, p.22).

Neste sentido, o conjunto de variáveis utilizadas no treino da rede SOM resultou da conjugação do "conhecimento apriorístico" sobre o tema e da necessidade de transformação das variáveis categóricas presentes na base dados original, uma vez que

os processos de optimização em distâncias preferivelmente utilizam variáveis de *input* intervalares.[88]

3.2.1. Análise descritiva das variáveis relevantes para a análise do problema

Como vimos, o HFT é resultado de um conjunto de condicionantes resultantes dos actuais sistemas electrónicos de negociação. Ou seja, de um ambiente exclusivamente automático de negociação onde a intervenção humana ocorre, na maior parte dos casos, somente no desenho de estratégias que são transpostas para os ditos algoritmos matemáticos, os quais identificam momentos específicos de mercado. As máquinas conectadas aos sistemas de negociação recebem e processam constantemente informações provenientes dos mercados, e, a uma velocidade extremamente alta, decidem o que e como comprar e vender.

A literatura sobre o tema é conscienciosa, na sua maioria, quanto aos critérios de identificação do "ambiente" HFT na negociação de um qualquer instrumento financeiro. Na identificação das variáveis importantes à modelação, procurou-se traduzir a maioria destes critérios, tendo seleccionado os seguintes:

A. Elevado rácio *Order-to-book* - Rácio obtido entre o número de ofertas transmitidas ao sistema de negociação e número de negócios realizados - o mesmo que dizer ofertas executadas.
B. Constante alteração da situação (estado da oferta no sistema de negociação) das ofertas, traduzido por um número significativo de cancelamentos/modificações das mesmas (ofertas).
C. Explosão de ofertas – muitas ofertas inseridas num pequeno intervalo de tempo.
D. Ofertas executadas num pequeno intervalo de tempo (milésimos de segundo).

---

[88] "…, as variáveis intervalares são as mais adequadas para servirem de base à analise de clusters". ….."As medidas geométricas baseadas nos espaços Euclidianos têm dominado a análise das relações de semelhança. Estas distâncias representam os objectos como pontos num qualquer espaço multidimensional, de forma a que as dissemelhanças observadas entre os objectos correspondam a distâncias métricas entre os respectivos pontos. Assim, a aplicação de metodologias de *clustering* passa na maioria das vezes pela utilização de índices de semelhança que respeitem as propriedades métricas." (Bação, cap. III, p. 12)

E. Ofertas identificadas como sendo realizadas para carteira própria (*proprietary trading*).
F. Outras variáveis importantes, como sejam: a quantidade de acções por oferta (moda); ofertas com preço definido; ofertas inseridas com data de validade somente para a sessão de bolsa;

Importa referir neste ponto que apesar de existir um consenso quanto à definição do HFT, quer as plataformas de negociação quer os reguladores europeus não dispõem de ferramentas para os identificarem. Ou seja, à excepção dos IF que actuam exclusivamente em nome próprio (*proprietary trading*), muito comum na realidade do mercado norte-americano, não se conhece, até o momento, uma forma fácil e expedita de os identificar, até mesmo porque com a cada vez maior capacidade de processamento das máquinas, maior é o detalhe das estratégias adoptadas pelos HFTr. Acresce a este facto o crescente volume de dados (ofertas e negócios) originado pela negociação.

Na análise aos dados obtidos procurou-se identificar o conjunto de informação que permitisse relevar os momentos *supra* referidos.

As variáveis identificadas *infra* (3.2.1.1.) reflectem ainda dois conceitos importantes para o estudo. O primeiro é o conceito de "latência", que se refere ao tempo de processamento/execução de uma qualquer oferta, este que a não existir identificará a não ocorrência, ou mesmo a inexistência de situações semelhantes ao HFT. Neste sentido, a confrontação da informação existente em cada variável teve em conta, na sua maioria, o tempo de processamento ao milissegundo. O segundo conceito que se procurou obter (ou reflectir) nas variáveis é o de "alta frequência", ou seja, o da concentração de um número significativo de ofertas processadas num reduzido intervalo de tempo. De forma a relevar estes momentos da negociação, a informação constante em cada registo, e para a quase totalidade das novas variáveis[89], reflecte a média do valor obtido a cada intervalo de tempo igual a 5 minutos de negociação. Para tal, subdividiu-se o normal período de negociação[90] em intervalos iguais a 5 minutos cada,

[89] Excepto as variáveis "A_RacioOrd", "C_OrdTime", "C_OrdTimeM", "E_LQtdOrd", "E_HQtdOrd" e "E_NQtdOrd".
[90] Período de negociação em contínuo 9:30:00-17:30:00. Ver anexo ao Manual de Negociação em http://europeanequities.nyx.com/sites/europeanequities.nyx.com/files/27102011_appendix_4-01.pdf

ou seja em 104 intervalos iguais[91]. A expressão *infra* identifica o cálculo realizado com base em cada um destes intervalos.

$$m = \frac{\sum_{i=1}^{N} \left(\frac{f_i}{F_i}\right)^2}{N}$$

onde **m** é o valor médio do quadrado dos pesos, obtidos em cada intervalo de tempo igual a 5 minutos, durante a sessão de bolsa.

Na expressão f_i identifica a frequencia das ofertas registadas pelo IF e processadas ao milésimo de segundo, enquanto F_i identifica o total das ofertas realizados pelo mercado (a totalidade dos IF) no mesmo intervalo temporal. *N* representa o número total de intervalos, igual a 104.

Na construção deste indicador utilizou-se como referência o índice de *Herfindahl–Hirschman* (HHI)[92], que procura medir o tamanho (peso ou relevância) de uma empresa em um determinado sector (industrial), relevando o impacto do seu *market share*, ou as alterações que deste decorrem, face ao demais concorrente. Por outras palavras, a concentração existente em um dado mercado.

Um outro objectivo, o de relevar a participação de um IF num dado contexto de negociação, é conseguido através do rácio entre as ofertas inseridas por estes (IF) em cada milésimo de segundo e o restante mercado, em cada intervalo temporal.

[91] Estes períodos identificam exclusivamente a negociação em contínuo.

[92] Vide, "The Herfindahl-Hirschman Index". Recuperado em 08 de Novembro, 2011. De http://www.justice.gov/atr/public/testimony/hhi.htm.

3.2.1.1.Procedimento para identificação das novas variáveis

A. *Order-to-book* - Rácio entre o número de ofertas e negócios é elevado.

Ofertas executadas num pequeno intervalo de tempo (milésimos de segundos)

Quadro 3.3 – *Variáveis produzidas para modelação – tipo A*

Variáveis	Descrição
A_RacioOrd	Rácio entre o nº total de ofertas introduzidas pelo IF numa dada sessão de bolsa para um instrumento financeiro face ao nº de negócios realizados (ofertas executadas) pelo mesmo. *Variável original: Ord_Stat*
A_OrdExec	Valor médio do peso relativo das ofertas executadas no milissegundo face ao número total de ofertas registadas pelo IF ("2"-Filled). *Variável original: Ord_Stat*
A_OrdExecM	Valor médio do peso relativo das ofertas executadas no milissegundo face ao número total de ofertas registadas na sessão de bolsa ("2"-Filled). *Variável original: Ord_Stat*

B. Constante alteração da situação das ofertas traduzida pela significativa quantidade de ofertas canceladas e/ou modificadas.

Quadro 3.4 – *Variáveis produzidas para modelação – tipo B*

Variáveis	Descrição
B_OrdMod	Valor médio do peso relativo das ofertas canceladas/modificadas ("4"-Cancelled / "5"-Replaced) no milissegundo face ao número total de ofertas registadas pelo IF. *Variável original: Ord_Stat*
B_OrdModM	Valor médio do peso relativo das ofertas canceladas/modificadas ("4"-Cancelled / "5"-Replaced) no milissegundo face ao número total de ofertas registadas na sessão de bolsa ("2"-Filled). *Variável original: Ord_Stat*

C. Explosão de ofertas – muitas ofertas num pequeno intervalo de tempo.

Quadro 3.5 – *Variáveis produzidas para modelação – tipo C*

Variáveis	Descrição
C_OrdTime	Valor médio do peso relativo das ofertas registadas pelo IF face ao número total de ofertas registadas na sessão de bolsa.
C_OrdTimeM	Valor médio do peso relativo das ofertas registadas pelo IF face ao número total de ofertas registadas na sessão de bolsa, com eliminação da sazonalidade na negociação (Não considera os 30 minutos iniciais e finais da sessão de bolsa, sendo analisado o período compreendido entre as 9:30 e as 17:00 CET)[93].

D. Ofertas identificadas como realizadas para carteira própria (*proprietary trading*).

Quadro 3.6 – *Variáveis produzidas para modelação – tipo D*

Variáveis	Descrição
D_OrdOri	Valor médio do peso relativo das ofertas processadas no milissegundo face ao número total de ofertas registadas pelo IF e com origem definida ("2"-House). *Variável original: Ord_Ori.*
D_OrdOriM	Valor médio do peso relativo das ofertas processadas no milissegundo face ao número total de ofertas registadas na sessão de bolsa e com origem definida ("2"-House). *Variável original: Ord_Ori.*

[93] Francisco,P. (2008), em estudo realizado sobre a dinâmica do livro de ofertas dos mercados Euronext, conclui que a "negociação está concentrada no início e no final da sessão exibindo um formato semelhante a uma "lua em quarto crescente".

E. Outras variáveis relevantes

Quadro 3.7 – *Variáveis produzidas para modelação – tipo E*

Variáveis	Descrição
E_ValTyp	Valor médio do peso relativo das ofertas processadas no milissegundo face ao número total de ofertas registadas pelo IF e com validade definida ("0"-Day). *Variável orginal: Val_Typ.*
E_ValTypM	Valor médio do peso relativo das ofertas processadas no milissegundo face ao número total de ofertas registadas na sessão de bolsa e com validade definida ("0"-Day). *Variável orginal: Val_Typ.*
E_OrdTyp	Valor médio do peso relativo das ofertas processadas no milissegundo face ao número total de ofertas registadas pelo IF e com preço definido ("2"-Limit). *Variável original: Ord_Typ.*
E_OrdTypM	Valor médio do peso relativo das ofertas processadas no milissegundo face ao número total de ofertas registadas na sessão de bolsa e com preço definido ("2"-Limit). *Variável original: Ord_Typ.*
E_LQtdOrd	Valor mínimo (*low*) relativo a Moda da quantidade de acções registada por oferta inserida pelo IF para o instrumento financeiro numa dada sessão de bolsa (Considera as ofertas de compras e as de vendas). *Variável original: Qtd_Ord.* **Nota**: pretende também verificar se o tamanho médio da oferta é reduzido.
E_HQtdOrd	Valor máximo (*high*) relativo a Moda da quantidade de acções registada por oferta inserida pelo IF para o instrumento financeiro numa dada sessão de bolsa (Considera as ofertas de compras e as de vendas). *Variável original: Qtd_Ord.* Nota: pretende também verificar se o tamanho médio da oferta é elevado.
E_NQtdOrd	Numero total de ofertas correspondente a moda. *Variável original: Ord_id.*

O resultado final é um conjunto de 16 variáveis (intervalares), calculadas para cada uma das seis acções seleccionadas.

3.2.2. Normalização dos dados

Neste trabalho o objectivo primeiro é identificar a actuação do HFTr, caracterizado por uma participação intensa e num pequeno intervalo de tempo, originando vários negócios, normalmente de quantidades reduzidas, realizadas para carteira própria, mas com um elevado volume de ofertas canceladas/modificadas. Importa pois que a informação reflectida em cada variável seja a apresentada à rede SOM nas melhores condições possíveis.

Verificou-se na análise exploratória dos dados que as variáveis apresentavam ordens de grandeza e unidades diferentes, bem como uma percentagem significativa de valores omissos, pelo que normalizar os dados se tornou um imperativo à boa prossecução do trabalho. Outra razão que obrigou à normalização deveu-se ao facto dos modelos não paramétricos[94], como o caso do SOM, assumirem que "as diferentes direcções do espaço de *input* possuem o mesmo peso" (Bação, 2007, cap.III. p.28).

Relativamente aos valores omissos, optámos por substituí-los pelo valor zero. Quanto à normalização, optou-se por utilizar, num primeiro momento, a normalização *zscore*, na qual cada variável de *input* é transformada por forma a ter média igual a zero e desvio padrão igual a um, uma vez que esta é a normalização efectuada, por defeito, pelo *software SAS Enterprise Miner*.

"A normalização pelo *zscore* transforma os valores da variável de *input*, de forma a que a sua média seja 0 e a variância 1. O primeiro passo consiste no cálculo da média e desvio-padrão dos dados de *input*. A seguir subtrai-se a cada *input* o valor da média e divide-se pelo desvio padrão" (Bação, 2007, cap. III, p.30). A fórmula da normalização *zscore* é a seguinte:

$$y' = \frac{y - média}{DP}$$ onde y é o valor original, y' o novo valor.

[94] Modelos que dependem essencialmente do uso das distâncias entre os dados.

No entanto os primeiros treinos realizados, demonstraram que a rede SOM considerava, de modo significato, alguns dos *outliers* da série de dados. Como esta informação (*outliers*), no caso estudado, é relevante para a modelação, optámos por recalcular as variáveis, desta feita aplicando a normalização sigmoidal.

"A normalização sigmoidal transforma os dados de *input* de forma não-linear para uma escala entre -1 e 1, utilizando a função sigmoidal. Começa por calcular a média e o desvio-padrão dos dados de *input*. Valores que se encontram a menos de um desvio-padrão da média são transformados de forma quase linear - caem na região linear da função sigmóide. Os *outliers* são comprimidos nas caudas da distribuição". (Bação, 2007, ibid., pág.30).

A fórmula é a seguinte:

$$y' = \frac{1 - e^{-\alpha}}{1 + e^{-\alpha}}$$

onde α corresponde ao *zscore*.

"A normalização pela função sigmoidal é particularmente apropriada quando existem *outliers* que desejamos incluir no conjunto de dados. Evita que os valores na amplitude média da distribuição sejam comprimidos para valores muito idênticos, sem perder a capacidade de representar *outliers* com valores muito elevados". (Bação, id.,Ibid., p.30).

3.3.Base de dados de treino e validação.

"A abundância de dados permite fazer face e minorar uma série de problemas. Quantos mais dados temos maior é a probabilidade de serem representativos dos casos que o modelo encontrará após a fase de treino. Quantos mais dados houver, mais fácil é para o modelo distinguir entre "ruído" e as verdadeiras relações existentes nos dados". (Bação, cap.II , p.33). " … a utilização de dados recentes é tanto mais importante quanto maior for a dinâmica do fenómeno a modelar." (Bação,ibid.,p.33)

Face à característica específica da informação, tendo em conta o objectivo do estudo e ainda o universo de dados disponíveis - informação relativa a seis diferentes acções - entendemos seleccionar para constituir a nossa base de dados de treino a

informação relativa ao instrumento financeiro identificado como "A", uma vez que este constitui o grupo de dados com maior frequência (4.388 registos).

Como refere Bação (2007), "se no futuro todos os vectores de *input* se apresentarem próximos dos exemplos utilizados durante a fase de treino, provavelmente o modelo apresentará um comportamento estável e com pequenos efeitos negativos. No entanto, se os vectores a classificar forem muito diferentes (ocupam áreas distantes no espaço de *input*) dos apresentados durante a fase de treino, então os resultados do modelo degradar-se-ão de forma acentuada. Em geral, quanto mais dados houver disponíveis para treinar, maior a probabilidade de construir um modelo robusto e estável" .

Neste sentido, entendemos utilizar a informação relativa às demais cinco acções ("B", "E", "I", "J" e "M") como base de dados de validação do modelo encontrado. No entanto, cada grupo de valores foi testado individualmente, o que significa dizer que foram antes utilizadas cinco base de dados diferentes para a validação. Este conjunto de dados é constituído, respectivamente, por 3.484, 3.235, 2.651, 2.897 e 3.086 registos.

A título informativo a *Figura 3.3. infra* apresenta a correspondência entre o real universo de dados (cerca de 9,7 milhões de ofertas), resultantes do ficheiro de negociação da Euronext e o universo de dados obtidos (cerca de 19,7 mil registos) após a produção das novas variáveis apresentadas à rede SOM.

	Intrumentos financeiros (acções)						
	A	**B**	**E**	**I**	**J**	**M**	**Total**
Ofertas	4.220.953	1.335.241	1.265.391	915.817	604.756	1.346.198	9.688.356
***Input* SOM**	4.388	3.484	3.235	2.651	2.897	3.086	19.741
Nº Sessões de Bolsa	86	85	85	86	85	76	

Figura 3.3 – Correspondência entre dados número de ofertas originais e os registos apresentados à rede SOM.

Fonte: informação produzida com base no ficheiro de ofertas.

3.4.Ferramentas analíticas.

No desenvolvimento das diferentes fases deste trabalho, recorreu-se às seguintes soluções informáticas para tratamento de dados primários – ofertas e negócios.

- *Software* Oracle para selecção dos dados relativos aos valores constituintes do indice PSI20;
- *Software* Microsoft Excel na agregação de informação entre os ficheiros de dados com ofertas e negócios (p.e. identificação da hora de execução da oferta e do efectivo tempo de processamento no milissegundo);
- *Software* Microsoft Access, na produção de novas variáveis;
- *Software SAS Enterprise Miner* como suporte à análise descritiva e na segmentação (nó de *Clustering* SOM) dos IF;

Neste terceiro capítulo é descrito a metodologia seguida na investigação do nível de utilização do conceito HFT no mercado de português de acções. São caracterizadas as bases de dados e realizada uma descrição das variáveis produdizadas, sendo referido as diferentes fases metodológicas seguidas. Por fim, são identificadas as aplicações informáticas que serão utilizadas.

O próximo capítulo apresenta os diferentes modelos utilizados para a melhor identificar/classificar a actuação dos IF na negociação das acções seleccionadas, bem como as suas características e os parâmetros utilizados na modelação.

4. ANÁLISE DE DADOS E RESULTADOS

Neste capítulo é efectuada uma breve análise à distribuição dos dados resultantes do processo de produção das novas variáveis e do procedimento de normalização sigmoidal.

No ponto seguinte, é iniciado o processo de segmentação dos dados com um procedimento de seleccção das variáveis relevantes ao treino do SOM, de entre o universo de (16) variáveis produzidas no capítulo anterior, as que melhor representam o objectivo pretendido.

O terceiro ponto apresenta a estrutura desenhada para o treino do SOM. Discute-se os resultados das primeiras especificações e a estrutura final da Base de Dados de Treino. Por fim é apresentada uma análise individual das variáveis e a sua importância na formação do *cluster*.

O quarto e último ponto apresenta o resultado da segmentação obtida com o modelo proposto no terceiro ponto e resultante do processo de validação da rede SOM.

4.1. Análise aos dados obtidos com as novas variáveis

Os gráficos *infra* identificam um exemplo do comportamento verificado para o registo da maioria das variáveis em estudo[95]. No exemplo, a variável "*A_RacioOrd*" [96], descreve os registos - resultantes da redução do espaço de *input* aquando da transformação das variáveis categóricas – obtidos para o instrumento financeiro "A" e relativos ao período em análise.

O exemplo permite identificar a transformação ocorrida na distribuição desta e demais variáveis após a normalização dos dados. A *Figura* 4.2 permite verificar que os valores encontram-se mais adequadamente distribuídos após a normalização.

[95] *Vide* em anexo (7.3) os gráficos de distribuição das variáveis seleccionadas para o treino do SOM.

[96] Rácio entre o número total de ofertas introduzidas por um IF numa dada sessão de bolsa e para um instrumento financeiro face ao número de negócios realizados pelo mesmo.

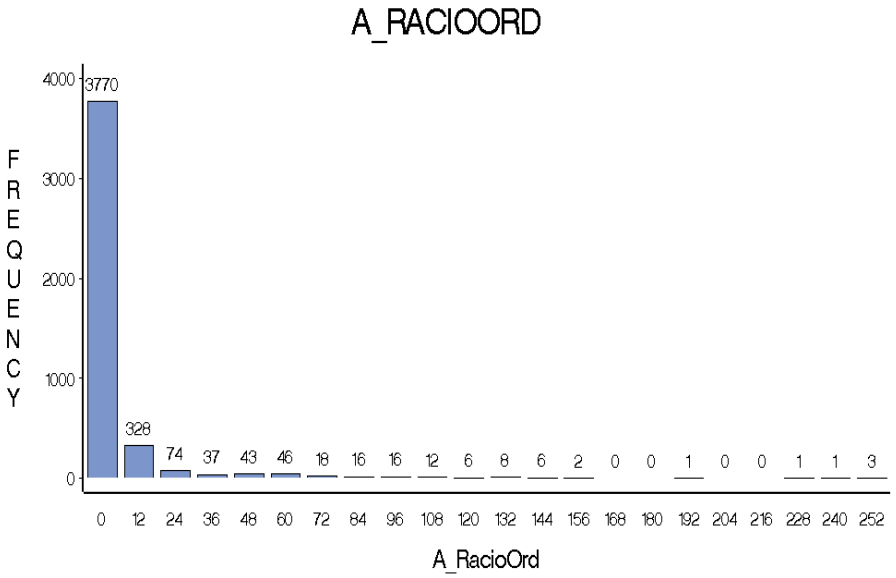


Figura 4.1 – Distribuição de frequência da variável "A_RacioOrd" – Agosto de 2011, antes da normalização sigmoidal.

Fonte: SAS Enterprise Miner – Multiplot.

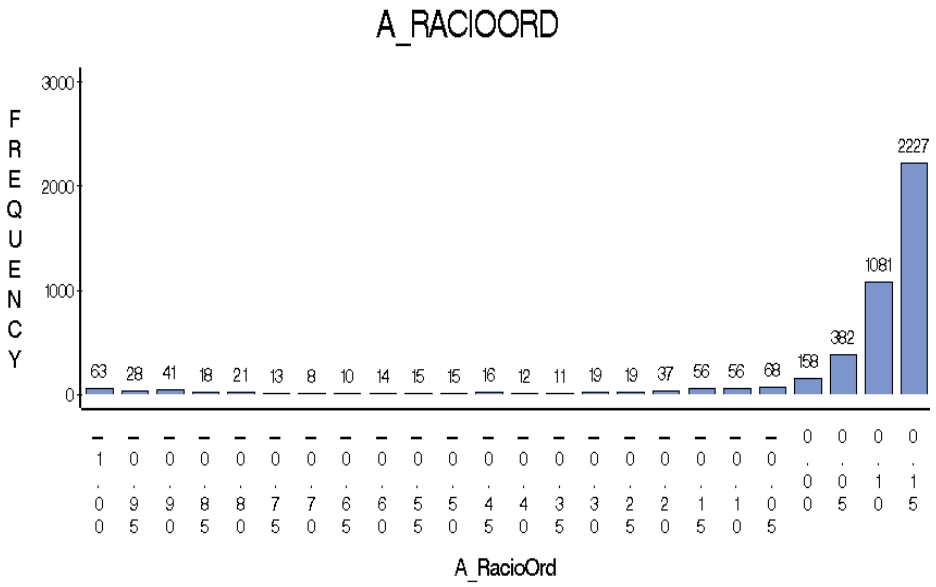


Figura 4.2 – Distribuição de frequência da variável "A_RacioOrd" - Agosto de 2011, após a normalização sigmoidal.

Fonte: SAS Enterprise Miner – Multiplot.

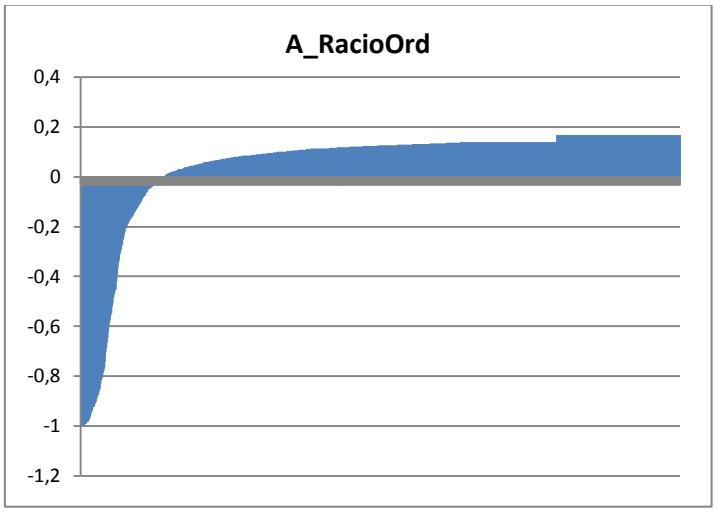


Figura 4.3 – Distribuição de frequência dos registos referentes a variável "A_RacioOrd" após normalização sigmoidal - Agosto de 2011.

Fonte: Informação produzida com base no ficheiro de ofertas e com recurso ao Microsoft Excel.

4.2. Classificação dos Intermediários Financeiros

O SOM é uma rede neuronal não supervisionada, "que pode ser encarada como uma projecção não linear de dados multidimensionais, estando por esta razão completamente livre para se ajustar aos dados de *input*. Por forma a processar os dados há que começar por ajustar os parâmetros, como o raio de vizinhança topológica, o número de neurónios, e a taxa de aprendizagem." (Loureiro e Bação, 2009).

A análise da segmentação produzida neste trabalho foi realizada em quatro etapas distintas, sendo as três primeiras um exercício de selecção/verificação das variáveis que melhor traduzem o objectivo prosseguido[97]. A primeira etapa visou avaliar o nível de classificação dos indivíduos (IF), segundo uma óptica de confrontação da sua actuação na "produção de informação"[98] no milissegundo face ao seu próprio desempenho na sessão de bolsa, e, a segunda face ao mercado – universo de ofertas inseridas para o instrumento financeiro numa dada sessão de bolsa. Posteriormente,

[97] *Vide* em anexo (7.4) os diferentes grupos de variáveis utilizadas nas referidas etapas.

[98] Ou seja, aquando do envio de ofertas para o sistema de negociação. Ofertas que foram processadas no milésimo de segundo. Uma oferta processada poderá ser equivalente a executada, cancelada ou modificada.

numa terceira etapa, utilizou-se uma combinação das duas primeiras. Para o cumprimento das primeiras etapas de identificação das variáveis relevantes ao treino do SOM[99], recorre-se a configuração padrão (*default*) do nó "SOM/Kohonen" do *SAS Enterprise Miner*.[100]

A análise de sensibilidade aos parâmetros do SOM foi realizada na quarta etapa. Nesta etapa, associa-se ao mapa topológico anterior um novo mapa obtido com base no nó "SOM/Kohonen" do *SAS Enterprise Miner*, sendo antes seleccionado o método "*Kohenen Self-Organizing Map"*, realizando-se vários testes à influência que os parâmetros reflectem no ajuste da rede aos dados.

O objectivo final da segmentação é conseguir identificar *clusters* onde as médias das variáveis seleccionadas identifiquem o máximo de dissemelhança possível da média da população. Constitui uma excepção, neste objectivo, os valores obtidos para as variáveis "E_HqtdOrd" e "A_OrdExecM" que pelo acima referido se pretende ser o mais parecida com a média da população.

A análise do agrupamento (em clusters) dos indivíduos em estudo - os IF intervenientes na negociação das acções seleccionadas – foi realizada com recurso aos *outputs* disponibilizados pela solução *SAS Interprise Miner*, como sejam, i) a matriz de distância *intra-cluster* visualizada no "*Multimensional scaling*"; ii) *Statistics* e iii) o ficheiro de *output* contendo a classificação dos registos (cluster) apresentados ao SOM.

4.3.Treino do SOM

No exercício de apresentação das variáveis ao SOM, realizado nas três primeiras etapas, verificou-se que o grupo de variáveis que melhor reflectia o comportamento pretendido de se verificar, resultou da relação dos indivíduos (IF) face ao mercado – 2ª etapa de processamento. As figuras *infra* identificam, respectivamente, as variáveis utilizadas para o treino da rede e a representação gráfica da distância intra-clusters, disponível no *software SAS Enterprise Miner*, o *multidimensional scaling*.[101]

---

[99] Variáveis identificadas no ponto "3.2.1.1. Prodedimentos para identificação das novas variáveis".

[100] A configuração *default* do *SAS Interprise Miner* utiliza um método "*Batch Self-Organizing Map*" com um mapa topológico de 4 linhas e 6 colunas (24 cluster).

[101] *Vide* anexo (7.5.1), a matriz de distância obtida com o treino de SOM para um mapa topológico de 4 linhas e 6 colunas.

Name	Model Role
DAT_IF	id
ORD_DAT	rejected
COD_INST	rejected
NUM_IF	rejected
A_RACIOORD	input
A_ORDEXECM	input
B_ORDMODM	input
C_ORDTIMEM	input
D_ORDORIM	input
E_ORDTYPM	input
E_VALTYPM	input
E_HQTD_ORD	input
E_NQTDORD	input

Figura 4.4 - Ficheiro de *input* (*Input Data Source*).

Fonte: SAS Enterprise Miner.

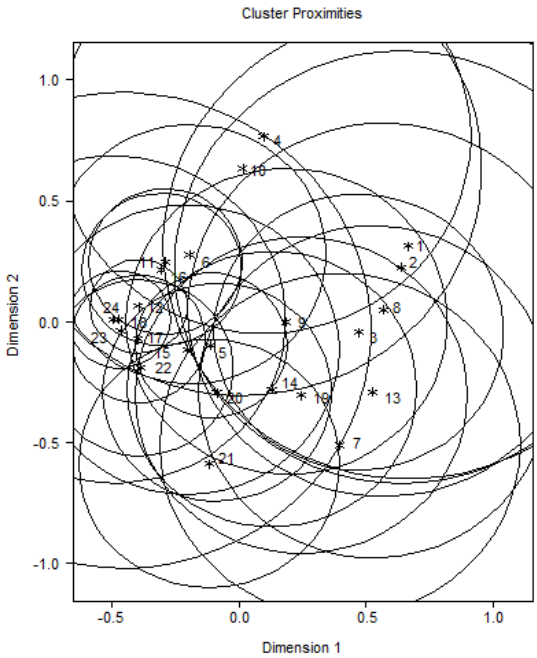


Figura 4.5 - Multidimensional scaling (rede SOM 4*6).

Fonte: SAS Enterprise Miner.

A decisão quanto ao número final de clusters é realizada na quarta etapa de processamento, e resultou da análise de sensibilidade aos parâmetros do SOM. Nesta fase, o valor de início da taxa de aprendizagem, o raio da função de vizinhança e o

número de épocas foi alterado conforme a alteração realizada à dimensão do mapa topológico, procurando com isto identificar a influência dos mesmos (parâmetros) na classificação dos indivíduos nos *clusters*[102].

Considerando que, a configuração padrão (*default*) do *SAS Enterprise Miner* (método "*Batch Self-Organizing Map*") conseguiu bem classificar algumas das variáveis em sete[103] dos 24 clusters produzidos, entendemos ligar ao referido mapa (*default*) o agora explorado[104]. A figura *infra* identifica a estrutura referida.

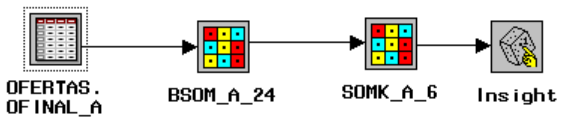


Figura 4.6 - Estrutura de Treino da rede SOM.
Fonte: SAS Entreprise Miner.

O objectivo desta conexão, entre diferentes mapas topológicos, é aproveitar a capacidade de discriminação conseguida com o exercício das fases anteriores. Verificada a maior, ou menor, influência de cada parâmetro, inclusive em mapas com diferentes dimensões (entre 4 e 12 clusters), procedemos ao treino final do SOM.

Os parâmetros finais utilizados para o treino do SOM equivalem a: i) uma taxa de início de aprendizagem 0.9 (0.01 final), ii) um raio de vizinhança incializado com um valor igual a 3 e iii) um número de épocas igual a 1000.

O mapa topológico escolhido, de superfície rectangular, apresenta uma dimensão de 2 linhas e 3 colunas (6 clusters). A figura *infra* identifica a estrutura do mapa escolhido.

---

[102] Este exercício de alteração foi efectuado repetidas vezes, no sentido de tentar refinar a capacidade de aprendizagem do algoritmo.

[103] *Vide* anexo (7.4) - gráficos de *output* do *SAS Entreprise Miner* com a comparação dos valores médios das variáveis entre os indivíduos da população para cada um dos sete *clusters* referidos.

[104] Mapa topológico (2 linhas e 3 colunas) obitido com base no nó "*SOM/Kohonen*" do *SAS Enterprise Miner*, com o método "*Kohenen Self-Organizing Map"*.

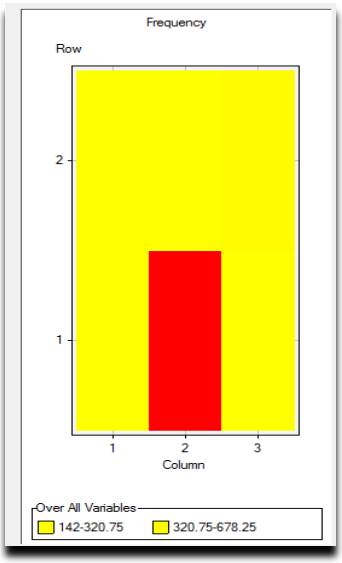


Figura 4.7 - Mapa topológico identificado para o treino do SOM.

Fonte: SAS Entreprise Miner.

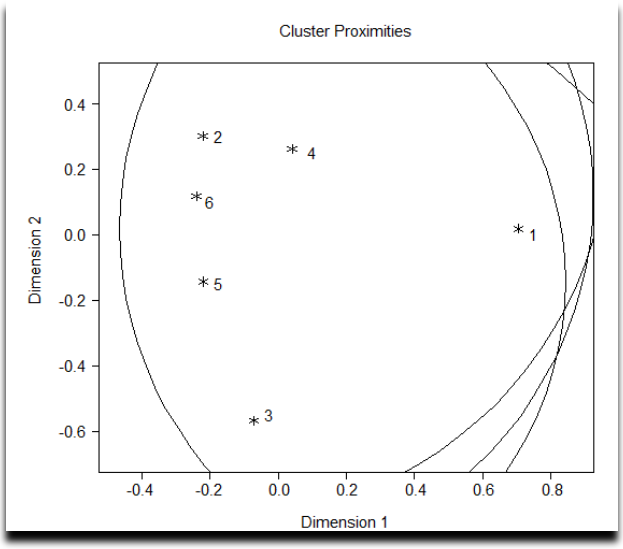


Figura 4.8 - Multidimensional scaling (rede SOM 2*3).

Fonte: SAS Enterprise Miner.

A *Figura 4.8* identifica o resultado do treino do SOM, onde se pode verificar um *cluster* (*cluster* 1) claramente diferenciado dos demais cinco clusters. A análise da

matriz de distâncias *infra* e do respectivo "erro de quantização" permitiu também verificar a qualidade do ajuste da rede neuronal aos dados.[105]

CLUSTER	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster
6	650	0.0532207664	1.1630325626	2	0.1917296639
2	3002	0.0385950172	1.1862889938	6	0.1917296639
5	211	0.0465099942	1.0606917464	6	0.2951968658
3	142	0.0735831361	1.3864744228	5	0.4234009456
4	207	0.2300188558	1.7485078009	2	0.7931719538
1	176	0.2307133024	1.1721339191	4	1.6581827171

Figura 4.9 – Matriz de distância do SOM.

Fonte: SAS Entreprise Miner.

Outra forma utilizada para corroborar o resultado da classificação passou pela análise dos valores médios das variáveis em cada *cluster*. Uma vez mais, verificou-se que as variáveis definidas atendem ao objectivo da classificação, produzindo *clusters* onde as médias das variáveis seleccionadas identificam o máximo de dissemelhança possível da média da população (p.e. o *cluster* 1) [106].

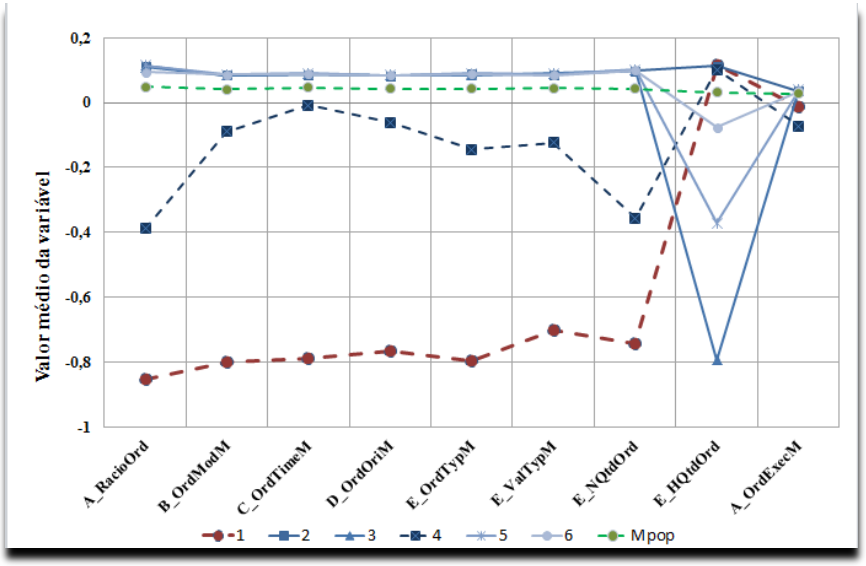


Figura 4.10 – Valor médio das variáveis nos Clusters

Fonte: Matriz de Distâncias SAS Enterprise Miner.

[105] O "erro de quantização" permite verificar a qualidade do ajuste da rede neuronal aos dados, este obtido através do "somatório das distâncias de cada indivíduo ao neurónio mais próximo" (Loureiro e Bação, 2009).

[106] *Vide* anexo (7.6.1) gráficos de *output* do *SAS Entreprise Miner* com a comparação dos valores médios das variáveis entre os indivíduos da população para o *cluster* 4.

Na figura *infra*, a avaliação do perfil do c*luster1* permite ainda identificar que as variáveis seleccionadas para o treino do SOM, permitiu classificar indivíduos com valores médios significativamente diferentes (dissemelhantes) dos valores médios obtidos para a população[107].

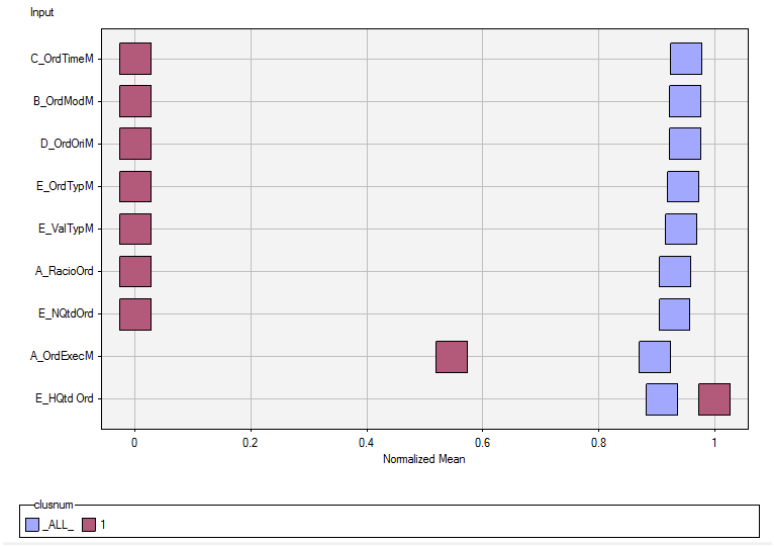


Figura 4.11 - Comparação entre valores médios das variáveis e os indivíduos da população.

Fonte: SAS Enterprise Miner.

Em resumo, o treino da rede é iniciado utilizando a configuração padrão do *SAS Enterprise Miner* para o nó "SOM/Kohenen". O método "*Batch Self-Organizing Map*", agrega os indivíduos (IF) apresentados à rede em 24 *clusters* distintos, onde sete destes (clusters) apresentam dissemelhanças expectáveis de se vir a encontrar com o treino do SOM. O treino prossegue com a associação de um segundo nó "SOM/Kohenen", no entanto o método seleccionado é o "*Kohenen Self-Organizing Map*". Neste novo mapa testamos a definição dos parâmetros "taxa de aprendizagem", "raio de vizinhança" e "número de épocas"[108]. Este passo é realizado algumas vezes procurando incitar a aprendizagem da

[107] *Vide* anexo (7.7) análise do ficheiro de *output* do SOM com o detalhe da classificação dos indivíduos para os demais cinco clusters.

[108] O aumento do número de épocas – de 1.000 até 3.000 – reflectiu exclusivamente e positivamente na média da variável "A_OrdExeM", que se pretende ser o mair semelhante possível da média da população, no entanto o erro de quantização obtido para o SOM foi maior. O número de indivíduos seleccionados foi o mesmo 176.

rede. Concluímos o mapa final com uma taxa inicial de aprendizagem de 0.9 (e final de 0.01), um raio de vizinhança menor (3) e um número de épocas igual (1000).

A análise da matriz de distância e do gráfico de médias - que permite a comparação entre valores médios das variáveis e os indivíduos da população - permitiu seleccionar o agrupamento de indivíduos (IF) com o maior grau de dissemelhança, objectivo pretendido com o treino da rede SOM.

Com base na correspondência dos indivíduos aos *clusters*, obtida como *output* do *software SAS Enterpise Miner*, verifica-se ainda que o *cluster* 1, apresenta a menor amplitude (máx-min) nas distâncias obtidas entre a posição dos padrões de *input* e a unidade onde se encontra mapeado. Também a distância *inter-clusters* bem define a sua diferença face aos demais cinco *clusters*.

Na agregação da informação obtida via referido ficheiro de *output*, identificámos seis indivíduos (IF)[109] seleccionados pelo SOM com um grau de semelhança significativa a uma actuação de HFTr. No entanto, somente os IF 66 e 101 apresentam uma frequência significativa para o período analisado. Os referidos IF foram seleccionados pelo SOM, respectivamente, em 98% e 90% das sessões em que participaram.

Número IF	% Freq face Sessões de bolsa
13	4,65%
15	2,33%
43	1,16%
66	**97,67%**
87	8,14%
101	**90,70%**

Figura 4.13 - Indivíduos seleccionados pela rede SOM na fase de treino.
Fonte: Estatística produzida com base no ficheiro de *output* do SOM.

[109] No ponto 3.1.2 identifica-se o universo de indíviduos (IF) intervenientes na negociação do instrumento financeiro "A" (igual a 71).

4.3.1. Análise individual das variáveis e a sua importância na formação do cluster.

De uma forma geral, na classificação o algoritmo atribuí "responsabilidades" às variáveis aquando da formação dos segmentos (clusters)[110]. Na *figura 4.11* pode-se verificar, para cada variável, a atribuição do nível de significância realizado pelo algoritmo aquando da criação do *cluster* 1. A influência da variável na determinação do segmento é identificada na ordenação (descendente) verificada no gráfico de distâncias entre as médias das variáveis e a população em estudo.

No quadro *infra* realiza-se um exercício de interpretação desta mesma atribuição, devendo a mesma se considerada em conjunto com a explicação apresentada no ponto 3.2.1.1. - Procedimentos para identificação das novas variáveis.

Quadro 4.1 – *Interpretação da relevância atribuída pelo SOM às variáveis.*

Variável	**Interpretação**
C_OrdTimeM	Identifica que o número médio de registos (ofertas), resultado da intervenção dos indivíduos (IF) do *cluster*, no conjunto das sessões de bolsa analisadas, foi em média significativamente superior a média do mercado. Identifica ainda, que os indivíduos seleccionados para o *cluster* actuam preferencialmente no intervalo temporal relativo a negociação em contínuo, uma vez que esta variável exclui a sazonalidade da primeira e última meia hora de negociação.
B_OrdModM	Identifica que a maioria das ofertas registadas pelos indivíduos do *cluster* sofreram cancelamento ou foram modificadas.
D_OrdOriM	Identifica que a maioria das ofertas foram registadas como sendo para a carteira própria.

(continua)

[110] O nível de significância de uma variável é definido pelo valor de cada vetor (relação dos seus vários registos) apresentado a rede. É por isto fundamental o processo de normalização das variáveis originais, principalmente aquando da existência de outliers, como no caso em estudo.

Quadro 4.1 – *Interpretação da relevância atribuída pelo SOM às variáveis.*

(continuação)

E_OrdTypM	Identifica que a maioria das ofertas registadas apresentava preço definido. Como referido, a utilização de soluções com base em algoritmos de negociação, não se coadunam com a tipologia de ofertas "ao mercado" (ao preço de mercado, sem preço definido), pelo elevado risco envolvido.
E_ValTypM	Identifica que a maioria das ofertas registadas pelos indivíduos do *cluster*, foram exclusivamente para uma dada sessão de bolsa. Um dos conceitos associados ao HFT é que as posições abertas durante uma sessão de bolsa são maioritariamente encerradas antes do fim do dia de negociação, o que exige que qualquer oferta transmitida ao sistema de negociação tenha validade definida, a sessão de bolsa.
A_RacioOrd	Identifica o rácio *order-to-book*, ou seja, a relação entre o número total de ofertas registadas para realização de um único negócio. Um rácio *order-to-book* elevado identifica, normalmente, a utilização, pelo indivíduo (IF), de soluções de negociação suportadas em algoritmos electrónicos, um dos critérios básicos à ocorrência de HFT.
E_NQtdOrd	Identifica a frequência relativa à moda da quantidade de cada oferta registada pelos indivíduos do *cluster*. A literatura sobre o tema (HFT) refere que nas estratégias de negociação implementadas pelos algoritmos electrónicos a quantidade das ofertas transmitidas ao sistema de negociação não se altera, sendo ainda na sua maioria de pequena dimensão. A selecção dos indivíduos (IF) do *cluster* aponta para uma elevada frequência.

(continua)

Quadro 4.1 – *Interpretação da relevância atribuída pelo SOM às variáveis.* (continuação)

A_OrdExecM	Identifica o peso relativo do número de ofertas executadas pelos indivíduos do *cluster* face ao mercado. A informação obtida com esta variável é complementar à variável "*A_RacioOrd*", não sendo no entanto de menor importância uma vez que considera as ofertas executadas no milissegundo, pelos indivíduos do *cluster* face à média do mercado. No exercício produzido pretende-se, ao contrario das demais, que esta variável apresente valores médios o mais próximo possível da média da população.
E_HQtdOrd	Identifica que a maioria das ofertas registadas pelos indivíduos do *cluster*, apresenta uma quantidade de pequena dimensão. Como referido, esta é também uma medida importante, pois associada à representatividade da variável "*E_NQtdOrd* " identifica claramente a utilização de algoritmos electrónicos pelos indivíduos do *cluster*. Como explicado, a quantidade reduzida das ofertas está associada ao controlo do risco envolvido em cada operação. Ou seja, o custo associado a uma operação será diminuto em face de uma estratégica mal conseguida, o que permite ao indivíduo (IF) reverter a posição então realizada, sem maiores prejuízos. Assim como com a variável "*A_OrdExecM*" pretende-se que esta variável apresente valores médios inferiores ou o mais próximo possível da média da população.

Pode-se referir que a atribuição de relevância realizada pelo SOM é bastante aceitável pois identifica como variável mais relevante a formação do cluster, uma variável ("*C_OrdTimeM"*) que reflecte a dimensão[111] da actuação dos indivíduos do *cluster* face ao mercado, seguido de um conjunto de (4) variáveis identificadoras, quer

[111] Medida pela frequência média de ofertas registadas no sistema de negociação.

de configurações utilizadas em algoritmos electrónicos, quer de actuação de um HFTr[112] e finaliza com as (3) variáveis que permitem corroborar situações de HFT e o cuidado na realização de operações com reduzido risco. A variável "*A_RacioOrd*" (rácio *order-to-book*) não sendo a mais importante na classificação do *cluster* é ainda significativa para corroborar o treino realizado.

4.4.Validação da rede SOM

Encontrado o mapa topológico adequado procedemos à realização da validação do SOM com base nas restantes base de dados, relativos às demais cinco acções identificadas, respectivamente, como "B", "E", "I", "J" e "M".

Como referido no ponto anterior, cada grupo de valores foi testado individualmente, o que significa dizer que foram antes realizados cinco exercícios de validação da rede. O conjunto de dados de validação é constituído, respectivamente por 3.484, 3.235, 2.651, 2.897 e 3.086 registos.

O exercício de validação da rede SOM, obtido com a fase de treino foi realizado com recurso ao nó "Score" do *SAS Enterprise Miner*, que permite aplicar sobre um novo conjunto de dados os parâmetros da rede (código gerado) obtidos através da configuração realizada na fase de treino. A figura *infra* identifica o modelo final.

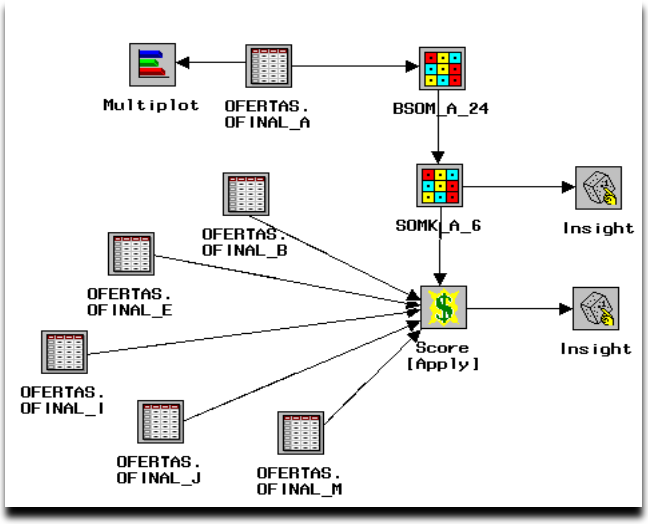


Figura 4.13- Estrutura utilizada para o treino e validação do SOM.

Fonte: SAS Entreprise Miner.

[112] Como sejam o elevado número de ofertas canceladas/modificadas, ofertas inseridas para uma única sessão de bolsa e com preço definido e identificadas como que realizada para carteira própria do IF.

Suscintamente o procedimento de *scoring* é realizado individualmente para cada base dado das cinco demais acções. Constituindo uma solução muito utilizada nos procedimentos de Datamining, o nó *Score* utiliza codigo SAS (“ SAS DATA step”), produzido com base no SOM, e aplicando-o posteriormente (no mesmo processamento) aos novos dados, então apresentados. A figura *infra* apresenta um resumo do exercício de validação realizado com recurso ao nó *Score* e resultante dos parâmetros obtidos com o treino da rede SOM[113].

Número Intermediário Financeiro	Intrumento Financeiro					
	A	B	E	I	J	M
	Frequência (%) participação face às Sessões de bolsa analisadas					
2				2,33%		
13	4,65%	**76,47%**	**96,47%**	**86,05%**	**94,12%**	**80,26%**
15	2,33%					
24		1,18%				
37		1,18%				
42						1,32%
43	1,16%					
44		1,18%	**12,94%**	8,14%	**11,76%**	**21,05%**
46					1,18%	
51						1,32%
55			1,18%			
56		2,35%				
63					4,71%	1,32%
66	**97,67%**				1,18%	
86					1,18%	
87	8,14%					
101	**90,70%**	**89,41%**	2,35%			**84,21%**
Nº Sessões de Bolsa	**86**	**85**	**85**	**86**	**85**	**76**

Figura 4.14 - Indivíduos seleccionados pelo SOM na fase de treino e validação.
Fonte: Estatística produzida com base nos ficheiros de output do SOM.

O SOM, classificou nesta fase, para o mesmo *cluster* (*cluster* 1), 11 novos individuos (IF), no entanto apenas um IF, codificado com o número 13, apresentou uma actuação significativa e equivalente aos IF 66 e 101 seleccionados na fase de treino. Os níveis de participação do IF 13, determinados pela frequência com que actua de forma

[113] O resultado da segmentação realizada na fase de validação é apresentado em (7.8) anexo.

semelhante a um HFTr, são superiores a 76% das 86 sessões de bolsa analisadas. Um quarto IF (44) apresentou também participações relevantes para quatro das seis acções analisadas, entre 8,14% e 21,05%.

No conjunto, o SOM seleccionou para o mesmo Cluster 17 IF, onde é clara a relevância do *modus operandi* dos indivíduos 13, 66 e 101 na negociação das acções seleccionadas para o estudo. No conjunto de indíviduos, poder-se-á atribuir ainda ao IF 44 uma relevância média, face à sua actuação em três dos valores analisados. Aos demais 13 IF atribuiu-se uma baixa relevância.

Este capítulo descreve o trabalho de selecção das variáveis ao treino da rede SOM, sendo definida a estrutura da rede que melhor identifica o grau de dissemelhança pretendido. Descreve ainda, o resultado do treino SOM e o processo de validação realizado com cinco diferente base de dados.

No próximo capítulo discutem-se os resultados e contribuições obtidas com o estudo, sendo ainda apresentado algumas propostas para investigação futura.

5. DISCUSSÃO

Neste estudo recorremos a capacidade de projecção não linear de dados multidimensionais da rede neuronal não supervisionada *kohonen Self-Organizing map* (SOM), com o objectivo de compreender em que medida, o HFT faz já parte, ou não, da realidade da negociação de acções no mercado de valores português.

A rede SOM treinada e posteriormente validada, analisou o padrão de actuação dos IF autorizados a actuarem no mercado Euronext Lisbon, com base no livro de ofertas do sistema de negociação da Euronext, procurando identificar se estas (ofertas) atribuem um grau de semelhança suficiente que permita equiparar os mesmos IF a uma actuação exclusiva de um HFTr.

As variáveis utilizadas para modelação do SOM foram criadas exclusivamente para este estudo, uma vez que as variáveis de interesse originais, resultantes do livro de ofertas são, na sua maioria, categóricas. As novas variáveis, intervalares, são mais adequadas para servirem de base à análise de segmentos (*clusters*). O objectivo de encontrar segmentos o mais dissemelhantes possíveis da população levou a considerar incluir no estudo um conjunto de *outliers*, facto que obrigou a uma normalização sigmoidal dos dados produzidos com as novas variáveis.

Na identificação das variáveis importantes à modelação procurámos traduzir a maioria dos critérios referidos na literatura sobre o "ambiente" HFT na negociação de um qualquer instrumento financeiro, como sejam o rácio *order-to-book*, o elevado registo de cancelamentos e/ou modificações das ofertas, os momentos de explosão de ofertas, o *proprietary-trading* e, principalmente, o conceito de "latência" e o processamento no milissegundo.

A base de dados primária, utilizada na produção das novas variáveis, agregou informação relativa a um conjunto de 12 acções constituintes do índice PSI20. Na análise da referida base de dados seleccionamos seis acções para as quais foi constituida uma segunda base de dados, desta feita para um período mais alargado (4 meses), e que contou com cerca de 9,7 milhões de registos. A base de dados de treino refere-se a um único título e a utilizada para a validação do modelo representa cinco outros. As referidas bases de dados (treino e validação) somatizam cerca de 19,7 mil registos.

A rede final, escolhida no decurso da fase de treino, é representada por um mapa topológico de 2 linhas e 3 colunas (6 clusters), utiliza o método "*Kohenen Self-Organizing Map*", e está configurada com uma taxa inicial de aprendizagem de 0.9 (e final de 0.01), um raio de vizinhança igual a 3 e um número de épocas igual a 1.000. Esta rede utiliza como *input* os dados de *output* de uma primeira rede SOM treinada com o método "*Batch Self-Organizing Map*".

No final da fase de treino, o SOM identificou num segmento único (Cluster1) um conjunto de seis IF com grau de semelhança suficiente para se equiparar à actuação de um HFTr. No entanto, apenas dois IF apresentavam relevância suficiente, uma vez que o algoritmo seleccionou os mesmos IF em mais de 90% das 86 sessões de bolsa analisadas. Os demais IF foram seleccionados em somente cerca de 10% das sessões.

Na fase de validação, apresentamos, para a mesma rede, informação relativa a cinco outras acções. No final do processo, o SOM seleccionou 11 novos IF não identificados na fase de treino. No entanto, e à semelhança do ocorrido na fase de treino somente outros dois IF apresentavam relevância sendo que apenas um deles apresentou níveis de participação superiores a 70%. No conjunto, o SOM seleccionou para o mesmo Cluster 17 IF, sendo que apenas três destes apresentam um nível significativo de relevância, que permita atribuir um grau de semelhança suficiente a equiparar a actuação de um *High Frequency Trader*.

Finalmente, podemos dizer que de forma genérica os resultados da análise correspondem às expectativas, uma vez que o SOM apresenta num único *cluster* um conjunto de IF com grau de semelhança suficiente à equiparação de actuação de um HFTr. Neste sentido, poder-se-á afirmar com este estudo que o mercado regulamentado de acções português é utilizado pelos IF numa lógica de negocição do tipo *high frequency trading*.

Contribuições

Uma das principais incógnitas, ainda por responder de forma clara e objectiva, refere-se ao grau de contribuição, ou não, da actuação de um HFTr na negociação de um qualquer instrumento financeiro. Neste sentido, o cada vez maior volume de dados na negociação, reflexo de todas as alterações referidas neste estudo mas principalmente da capacidade de processamento das máquinas, não facilita (ou contribui) às entidades reguladoras dos mesmos, o normal exercício de supervisão da negociação.

Verificámos com este estudo que a implementação de uma rede neuronal não supervisionada, do tipo *Kohenen Self-Organizing Map*, poderá contribuir à identificação, em um universo alargado de dados da negociação, dos IF que apresentam um grau suficiente de semelhança à actuação de um HFTr, constituindo assim um ponto de partida a uma investigação mais apurada da negociação de um qualquer instrumento financeiro.

Propostas para investigação futura

Um exercício de significativa relevância para a análise e compreensão do expectro de actuação dos HFTr será associar as descobertas proporcionadas por este trabalho ao impacto da volatilidade diárias das acções. Ou seja, o objectivo passaria por analisar em pormenor, tendo em conta o *ouput* produzido pelo SOM, o impacto de actuação do HFTr na formação dos preços em mercado, bem como da liquidez real/potencial associada ao mecanismo de formação de preço, sendo esta análise realizada ao nível das ofertas.

Um segunda proposta seria a aplicação do procedimento as restantes 14 acções do índice PSI20 durante período mais alagardo, por exemplo os últimos 5 anos. Este período temporal permitiria identificar, não só os IF com um grau de semelhança a actuação de um HFTr mas, também, o momento de mudança a estrutura do mercado português de acções, o que permitiria confrontar com as demais realidades europeias e mundias o tempo de reacção a adopção deste novo fenómeno mundial.

Por fim, importa referir que a evolução verificada em menos de uma década na negociação em mercado de valores mobiliários, resulta em grande medida, do

desenvolvimento tecnológico mas também de uma regulamentação adequada à introdução de novos produtos e procedimentos, vocacionados para o desenvolvimento económico das empresas e dos mercados e, em semelhante medida, com a segurança do investidor que nele actua. No entanto, a vertente regulamentar, que não fez parte do escopo deste estudo, poderia ser explorada em investigação futura. O objectivo principal seria identificar a mudança dos padrões de actuação dos IF em mercado de bolsa, aquando da alteração regulamentar.

6. BIBLIOGRAFIA

Almeida, R. (2011). *Classificação de churn no Seguro Automóvel*. (Dissertação de Mestrado, ISEGI, Lisboa). Recuperado em 26 de Abril de 2011, de http://dspace.fct.unl.pt/dspace/bitstream/10362/5389/1/TEGI0276.pdf

Aldridge, I. (2010). *High-Frequency Trading: A Practical Guide to Algorithmic Strategies and Trading Systems*. Haboken, New Jersey: John Wiley & Sons.

Avellaneda, M. & Stoikov, S. (2008). High-frequency trading in a limit order book. *Quantitative Finance, 8*(3), 217-224.

Bação, F. (2007). *Data Mining*. Unpublished manuscript. Lisboa.

Brogaard, J. A. (2010). High Frequency Trading and Its Impact on Market Quality. *5th Annual Conference on Empirical Legal Studies Paper*. Retrieved November 22, 2010 from: http://ssrn.com/abstract=164138768).

Beddington, J. at al. *The Future of Computer Trading in Financial Markets*. Foresight. Retrieved October 06, 2011 from: http://www.bis.gov.uk/assets/bispartners/foresight/docs/computer-trading/11-1276-the-future-of-computer-trading-in-financial-markets.

Comissão Europeia. (2006). *Directiva 2006/73/CE da Comissão, de 10 de Agosto de 2006, que aplica a Directiva 2004/39/CE do Parlamento Europeu e do Conselho no que diz respeito aos requisitos em matéria de organização e às condições de exercício da actividade das empresas de investimento e aos conceitos definidos para efeitos da referida directiva*. Jornal Oficial da União Europeia. Publicado 02 de Setembro de 2006. L 241, pp. 26-58. Recuperado em 08 de Abril de 2011, de http://eur-lex.europa.eu/LexUriServ/site/pt/oj/2006/l_241/l_24120060902pt00260058.pdf

Comissão Europeia. (2004). *Directiva 2004/39/CE do Parlamento Europeu e do Conselho, de 21 de Abril de 2004, relativa aos mercados de instrumentos financeiros. Altera as Directivas 85/611/CEE e 93/6/CEE do Conselho e a Directiva 2000/12/CE do Parlamento Europeu e do Conselho e revoga a Directiva 93/22/CEE do Conselho*. Jornal Oficial da União Europeia. Publicado 30 de Abril de 2004. L 145, pp. 1,44.

Recuperado em 08 de Abril de 2011, de http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=oj:l:2004:145:0001:0044:pt:pdf

European Commission (2010). *Public Consultation - Consultation Review of the Markets in Financial Instruments Directive (MIFID).* European Commission, pp. 1, 83. Retrieved March 29, 2011 from http://ec.europa.eu/internal_market/consultations/docs/2010/mifid/consultation_paper_en.pdf

Deboeck, G. & Kohonen, T. (1998). *Visual Explorations in Finance with Self-Organizing Maps*. New York: Springer.

Francisco, P. (2008). *Euronext Stock Exchange Order Book and Order Flow Dynamics.* (Dissertação MBA, ISEG, Lisboa). Recuperado em 05 de Maio de 2011, de https://aquila.iseg.utl.pt/aquila/instituicao/ISEG/lateral/investigacao/dissertacoes/mestrados/gestao---mba/2008

Hasbrouck, J. & Saar, G. (2011). *Low-Latency Trading*. (Nº. 35-2010). Ithaca, NY: Johnson School. Retrieved February 1, 2011 from: http://ssrn.com/abstract=1695460.

Hendershott, T., Jones, Charles M. & Menkveld, A.J., (2010). *Does Algorithmic Trading Improve Liquidity?*. Journal of Finance, Vol. 66, pages: 1-33.

Hendershott, T. & Moulton, P. (2009). *Speed and stock market quality: The NYSE's Hybrid*. Berkeley: University of California. Retrieved February 02, 2011 from: http://ssrn.com/abstract=1159773.

Horta, P. e Mendes, C. (2007). *Aplicação das Redes Neuronais Artificiais à Detecção dos Mercados Euronext Mais Rentáveis*. CEFAGE-UE. Recuperado em 04 de Abril de 2011, de http://econpapers.repec.org/paper/cfewpcefa/2007_5f05.htm.

Kiang, M., Hu, M., & Fisher, D. (2006). An extended self-organizing map network for market segmentation— a telecommunication example. *Decision Support Systems*, *42*(1), pages. 36-47.

Lee, K., Booth, D., & Alam, P. (2004). *Using an Extended Self-Organizing Map Network to Forecast Market Segment Membership*. In P. G. Zhang (Ed.), Neural Networks in Business Forecasting (pp. 142-157). Hershey, USA: IRM Press.

Loureiro, M., Bação, F. (2009). *O Self-Organizing Map como Ferramenta na análise geo-demográfica.* Instituto Superior de Estatística e Gestão de Informação Universidade Nova de Lisboa. Recuperado em 02 de Novembro de 2011, de http://www.apgeo.pt/files/docs/CD_V_Congresso_APG/web/_pdf/C18_Out_Loureiro_Bacao.pdf

McNelis, P. D. (2005). *Neural networks in finance: gaining predictive edge in the market*. Burlington, USA: Elsevier Academic Press.

Menkveld, A. J. (2011). *High Frequency Trading and The New Market Makers*. Retrieved March 15, 2011 from http://ssrn.com/abstract=1722924.

Netherlands Authority for the Financial Markets (AFM). (2010). *High Frequency Trading: The application of advanced trading technology in the European marketplace*. Retrieved April 15, 2011 from http://www.afm.nl/layouts/afm/default.aspx~/media/files/rapport/2010/hft-report-engels.ashx

Neto, M. C. (2008). *Séries Exógenas e Combinação de Redes Neurais aplicada ao mercado financeiro*. (Dissertação de Mestrado, UFPE, Pernanbuco). *ufpe.br*. Recuperado em 23 de Maio de 2011, de http://www.cin.ufpe.br/~viisar/doc/dissertations/2008_msc_mcan.pdf.

Nyse Euronext. (2011). Anexo a Instrução 4-01 - Euronext Cash Market Trading Manual. Actualizado em 26 de Outubro de 2011. Recuperado em 17 de Novembro de 2011, de http://europeanequities.nyx.com/sites/europeanequities.nyx.com/files/27102011_appendix_4-01.pdf

P. Ravi Kumara & V. Ravi (2006). Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. *European Journal of Operational Research*, Volume 180, Issue 1, 1 July 2007, Pages 1-28

Riordan, R. & Storkenmaier, A. (2008). Latency, Liquidity and Price Discovery. *16th Annual Meeting of the German Finance Association 2009 Paper*. Retrieved March 29, 2011 from: http://ssrn.com/abstract=1247482.

7. ANEXOS

7.1. Conceitos

O desenvolvimento tecnológico verificado ao longo dos anos, levou a maioria das estruturas de negociação mundiais a adoptarem os sistemas de negociação electrónicos (ET). Na sua concepção, suportada em potentes computadores, a "lógica de negócio" é traduzida e incorporada num conjunto significativo de programas electrónicos (PT) que recorrem a parâmetros prévia e regularmentarmente definidos[114].

No contexto europeu, o fomento à concorrência entre as estruturas de negociação, implementado pela DMIF em finais de 2007, intensificou o recurso à utilização de alta tecnologia, o que permitiu melhorar a qualidade quer das estruturas de negociação quer dos meios de acesso (conexão) às mesmas, ambos um incentivo à adopção de estratégias de negociação suportadas em algoritmos electrónicos (AT).

Associado à caracterização da solução HFT encontramos um conjunto distinto de conceitos, que se pretende explicar, para melhor compreender a envolvente desta investigação, nomeadamente o ambiente propício à utilização da solução (HFT), as principais estratégias de negócio que recorrem à solução e o conceito de latência, quer dos sistemas de negociação quer dos acessos (conexão) às estruturas de negociação que os suportam. A definição cuidada de cada um destes conceitos constará, como referido, da primeira parte do trabalho de investigação. Nos pontos seguintes apenas apresentamos uma breve descrição de cada um destes conceitos.

[114] Regulamentação emitida pela entidade de supervisão e pelas sociedades gestoras que identificam a lógica e as directrizes do mercado e da negociação em bolsa. Em Portugal, respectivamente, a CMVM e a Euronext Lisbon;

7.1.1. High frequency trading.

O HFT é um método que utiliza tecnologia avançada (*software* inteligente) para implementar uma determinada estratégia de negociação. No entanto, o HFT não pode ser visto de forma separada como uma estratégia de negociação [115].

7.1.2. Algorithmic trading.

Algorithmic trading (algoritmos de negociação) podem ser definidos como "introduzir uma ordem de compra ou venda com uma quantidade específica (definida) em um modelo quantitativo, que gera automaticamente o timing de execução e a quantidade das ofertas (transmitidas para o sistema de negociação), suportado em objectivos especificados por parâmetros e limitações do algoritmo" .

As regras incorporadas ao modelo quantitativo determinam o tempo ideal para a execução/transmissão de uma ou mais ofertas aos sistemas de negociação, tendo em consideração um princípio básico que se traduz em causar o mínimo de impacto sobre o preço do instrumento financeiro. Outra solução de utilização destes algoritmos, agradável a um conjunto significativo de gestores de carteiras de investimento, é a facilidade com que o algoritmo digere as intenções de compra ou venda em ofertas de pequena dimensão. Por exemplo, em vez de enviar uma oferta para o sistema de 1.000.000 de acções do instrumento XPTO, uma estratégia de negociação algorítmica pode transmitir ofertas de 1.000 acções, ou menores, a cada 30 segundos, ao longo de várias horas ou durante toda a sessão de bolsa. Ao subdividir o volume da oferta em pequenas quantidades, o IF é capaz de "disfarçar" as suas ofertas e participar da negociação de um dado instrumento durante toda a sessão ou num menor intervalo de tempo. O prazo depende do objectivo e da estratégia definida, e quão agressivos (ou discretos) desejam ser na negociação de um qualquer instrumento financeiro.

---

[115] "A specific type of automated or algorithmic trading is known as high frequency trading (HFT). HFT is typically not a strategy in itself but the use of very sophisticated technology to implement traditional trading strategies." (Definição apresentada pela Comissão Europeia em documento de base a consulta pública de revisão da DMIF http://ec.europa.eu/internal_market/consultations/docs/2010/mifid/consultation_paper_en.pdf, pág.14);

7.1.3. Âmbito de actuação.

Estimativas do nível de participação do HFT no mercado europeu de instrumentos financeiros não são claras, no entanto sabe-se que o HFT representa uma parte considerável, e crescente, na maioria das estruturas de negociação europeias, presumindo-se ainda que este crescimento irá continuar por algum tempo. Segundo estudo recente da AFM (2010, p. 11), "os valores apresentados para o mercado europeu variam enormemente, entre 13% e 40% a 50% do volume de negócios" [116].

Esta limitada objectividade na definição do volume de negociação realizado com recurso ao HFT deve-se basicamente a duas questões principais. A primeira relaciona-se com a própria definição do fenómeno HFT, não existindo ainda um consenso sobre a terminologia. No entanto, quer a Comissão Europeia (CE) quer a ESMA encontram-se imbuídas em solucionar esta questão. Em ambos os organismos, grupos de trabalho foram formados para rever, entre outros, quer o conceito deste novo fenómeno de mercado quer todas as questões relacionadas com o mesmo, como por exemplo, a *co-location* (localização da infraestrutura do IF junto da infraestrutura do operador de mercado), o *sponser access* (acesso patrocinado), o *tick size* e o preçário cobrado pelas entidades gestoras de mercados (bolsas) no fomento ao desenvolvimento do fenómeno. A segunda questão, esta, objectivo principal deste estudo (definição do nível de participação do HFT no mercado português de valores mobiliários) está relacionada com a correcta identificação das operações realizadas com recurso ao HFT.

Como referido, o recurso à utilização de algoritmos electrónicos (ET) na negociação de instrumentos financeiros é não só uma solução há muito utilizada por grande parte dos intervenientes no mercado de instrumentos financeiros (e.g. Bancos, Grandes empresas de investimentos, Hedge Funds, entre outros) como também de difícil detecção, o que torna a sua estimativa, na maior parte das vezes, elevada.

---

[116] Informação resultante do resultado do *Call for evidence* CESR/10-142, do CESR, publicado em 04 de Maio de 2011. Recuperado de http://www.esma.europa.eu/index.php?page=responses&id=158.

7.1.4. Estratégias.

As estratégias de negociação utilizadas no HFT não são novas, e estão divididas em dois tipos diferentes: i) *market making* (*statistical arbritage*) e ii) *low latency*.

Market making: O objectivo desta estratégia, como o próprio nome indica, é criar liquidez na negociação dos instrumentos financeiros admitidos à negociação. Especificamente esta estratégia consiste em procurar preços para um instrumento financeiro em outra plataforma onde o mesmo esteja também admitido/seleccionado à negociação. O *spread*, diferença obtida entre os respectivos preços, é o ganho do *market making*. Esta estratégia obriga logicamente a que o High Frequency Trader (HFTr) esteja conectado às outras plataformas de negociação, onde o instrumento financeiro esteja admitido à negociação (eg. um MTF).

Com o advento da DMIF verificámos uma multiplicação de novas plataformas multilaterais de negociação (MTF) por toda a Europa (cerca de 142, segundo informação da base de dados da ESMA) o que facilita a aplicação de uma estratégia do tipo *market maker*. Os *market makers* providenciam liquidez a estas estruturas de negociação introduzindo ofertas (intenções de compra e venda) que pretendem negociar em outras estruturas, considerando um ligeiro *spread* (diferença de preços). Estas estruturas fomentam este exercício, reduzindo o custo por transacção ao HFtr. Preçarios "make-taker" são um exemplo utilizado por estas plataformas, onde uma oferta que permaneça no livro de ordens por algum tempo (actuação passiva) tem custo mais reduzido (ou até um benefício através e um *rebate*) do que uma oferta que seja introduzida e imediatamente executada (actuação agressiva).

Low latency strategies : O mais importante e um factor de sucesso deste tipo de estratégia, é ser mais rápido que o resto do mercado. Esta é uma categoria muito ampla composta de muitos tipos diferentes de estratégias. Estas estratégias consistem em ter o mais rápido dos sistemas e a melhor conexão com as estruturas de negociação:

a. Searching out limit orders;
b. Analysing the way;
c. By moving the market;
d. Building your own.

Os algoritmos que utilizam este tipo de estratégias (*low latency*) são conhecidos como algoritmos agressivos.

7.1.5. Latência.

Reduzir o tempo entre o momento em que o preço de uma oferta é "visualizado", ter em conta o momento do mercado e reagir confirmando/enviando uma oferta em sentido contrário, permitindo assim executar uma operação, é um dos principais objectivos para o sucesso do HFT. Neste contexto, o tempo de processamento (latência), agora medido em microssegundos, é o argumento (variável) principal no sucesso de implementação de estratégias de mercado, devendo ser analisado sob duas perspectivas distintas:

a. *Round-trip latency* – é o tempo que o sistema de negociação leva para receber (via *firewall*), executar, se for o caso, e enviar a resposta (via *firewall*) sobre uma oferta. Este tempo é medido entre o momento em que a informação chega e sai do sistema de negociação.
b. *Proprietary latency* – é o tempo obtido pela distância entre o IF e o sistema de negociação. Ao contrário do sistema de negociação, a qualidade das conexões utilizadas e a velocidade da sua estrutura computacional, bem como o tipo de algoritmo utilizado pelo IF podem alterar o tempo de processamento de uma oferta. O IF pode sempre optimizar este tempo melhorando a qualidade da sua estrutura (software e hardware) e as conexões à plataforma de negociação.

No contexto de participação do HFTr, obter uma menor latência (Proprietary latency) depende de múltiplos factores, nomeadamente, i) da sofisticação e da complexidade dos algoritmos utilizados; ii) da capacidade da sua infra-estrutura (IT System); iii) da "largura de banda" , rapidez e estabilidade da conexão com o sistema de negociação (*network latency*); iv) da topografia da sua estrutura, ou seja, o número de conexões (*routers*) utilizadas na comunicação com o sistema de negociação; v) da distância física entre o servidor que processa os algoritmos e o sistema de negociação (*propagation delay*) e não menos importante vi) se o IF actua directamente junto do sistema de negociação (DMA) ou utiliza um acesso intermediado (*sponsored access*).

Os desenvolvimentos tecnológicos têm vindo a permitir reduzir constantemente o tempo de processamento de uma oferta. A velocidade é agora expressa em

microssegundos (1 milionésimo do segundo) e a expectativa futura próxima será para o *nanosegundo*.

Logicamente que a importância da latência está relacionada com o tipo de estratégia utilizada pelo IF, enquanto esta estiver dentro de um intervalo de menos de cerca de um décimo do segundo.

7.1.6. Co-location (colocalização).

Uma forma de reduzir a latência é dispor o servidor utilizado para "correr" os algoritmos electrónicos próximo ao sistema de negociação. A esta proximidade do sistema de negociação electrónico (ET), denominou-se chamar "colocalização" (co-location). Face ao custo previamente acordado com as estruturas de negociação (Bolsas), estas oferecem aos seus Membros (IF) a oportunidade de arrendar um espaço[117] na mesma estrutura utilizada para suportar os servidores do sistema de negociação. A pequena distância entre as máquinas permite obter uma vantagem competitiva face aos restantes IF, uma vez que o tempo de resposta ao envio de uma oferta e recepção da informação do mercado é bastante reduzido (milissegundos).

Embora os recentes desenvolvimentos tecnológicos facilitem esta solução, o conceito de co-localização não é recente. Os profissionais do mercado sempre precisaram de estar próximos do local onde o preço é formado, o que permite reagirem mais rápido às alterações e às oportunidades do mercado. O conceito de co-localização é, portanto, comparável com o sistema utilizado nos recintos de negociação (*floor*), onde os IF detinham cabinas de negociação que permitiam aos seus operadores estar o mais próximo das estruturas das bolsas onde ser registavam as operações.

A co-localização, a par dos custos envolvidos na solução, oferece ao IF a oportunidade de estar "em pé de igualdade" com outros IF de grande dimensão, uma vez que a distância física à plataforma de negociação tem resultado directo na latência dos participantes do mercado.

Como a latência da estrutura de conexão do IF (*proprietary latency*) é também um factor preponderante nesta competição ao melhor momento de mercado, muitos IF

[117] Cabine (racks) onde são instalados os servidores utilizados pelos HTFr para armazenar e reproduzir os algoritmos de negociação (AT);

optam por localizar as suas estruturas de trabalho próximas à estrutura (edificação) das sociedades gestoras de mercados (bolsas)[118], adquirindo assim uma vantagem "arbitrária" sobre os IF demais participantes no mercado.

7.1.7. Sponsored access (acesso patrocinado).

Existem hoje disponíveis, aos participantes no mercado de instrumentos financeiros (IF), diferentes formas de conexão às estruturas de negociação (ET). A mais eficiente (rápida), e consequentemente a que envolve maiores restrições (exigências) e custos, é a ligação directa à estrutura de negociação, só permitida aos IF membros do mercado[119]. Existem no entanto, formas alternativas de acesso às estruturas de negociação, algumas destas especialmente atraentes para as estratégias desenvolvidas por HFTs, não membros, como sejam o DMA e o SA.

Os denominados acesso directo ao mercado (DMA) e "acesso patrocinado" (SA), consistem numa forma adaptada de conexão que permite a um cliente de um IF, membro de uma plataforma de negociação (bolsa), obter acesso directo ao sistema de negociação (ET), sem ter de se tornar um membro.

A diferença básica consiste em que, no DMA todo o fluxo de transacção efectuado pelo IF (cliente) passa através da estrutura de conexão do IF membro, de modo a que todos os mecanismos de controlo sejam automaticamente exercidos sobre as ordens transmitidas à estrutura de negociação (ET). No SA o IF membro apenas acompanha a actuação do IF junto da estrutura de negociação, recebendo a informação relativa à actuação do IF cliente. A par dos menores custos envolvidos, os ganhos obtidos com a velocidade na execução das suas ofertas e o efectivo anonimato, face ao mercado, tornam estas soluções bastante atractivas aos HFTr.

---

[118] A Nyse Euronext passou a disponibilizar este serviço ao IF europeus a partir de Abril de 2008;

[119] Membro é a designação atribuída ao IF autorizado a actuar numa determinada estrutura de negociação, esta de propriedade de uma entidade gestora de mercado ou sistema (eg. Euronext Lisbon);

7.2. Detalhe Tabelas de dados Originais

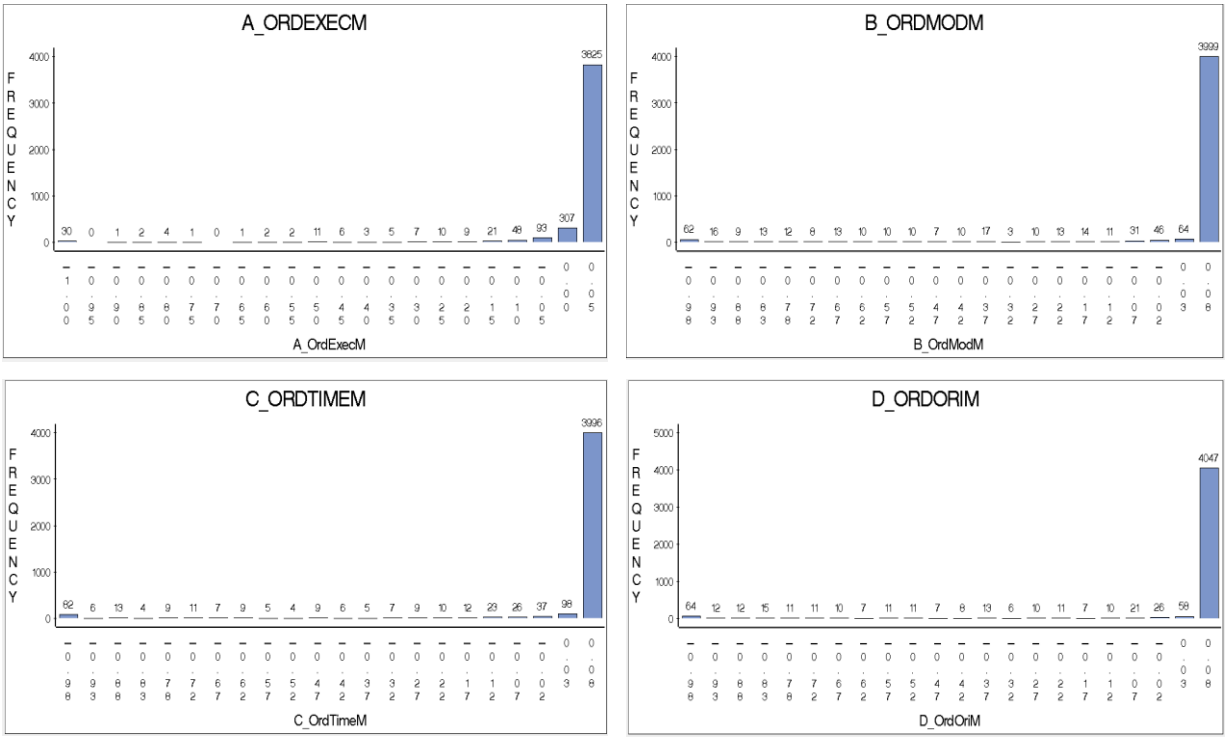
Quadro 3.1 – *Base de dados de ofertas -Variáveis Originais (Detalhe)*

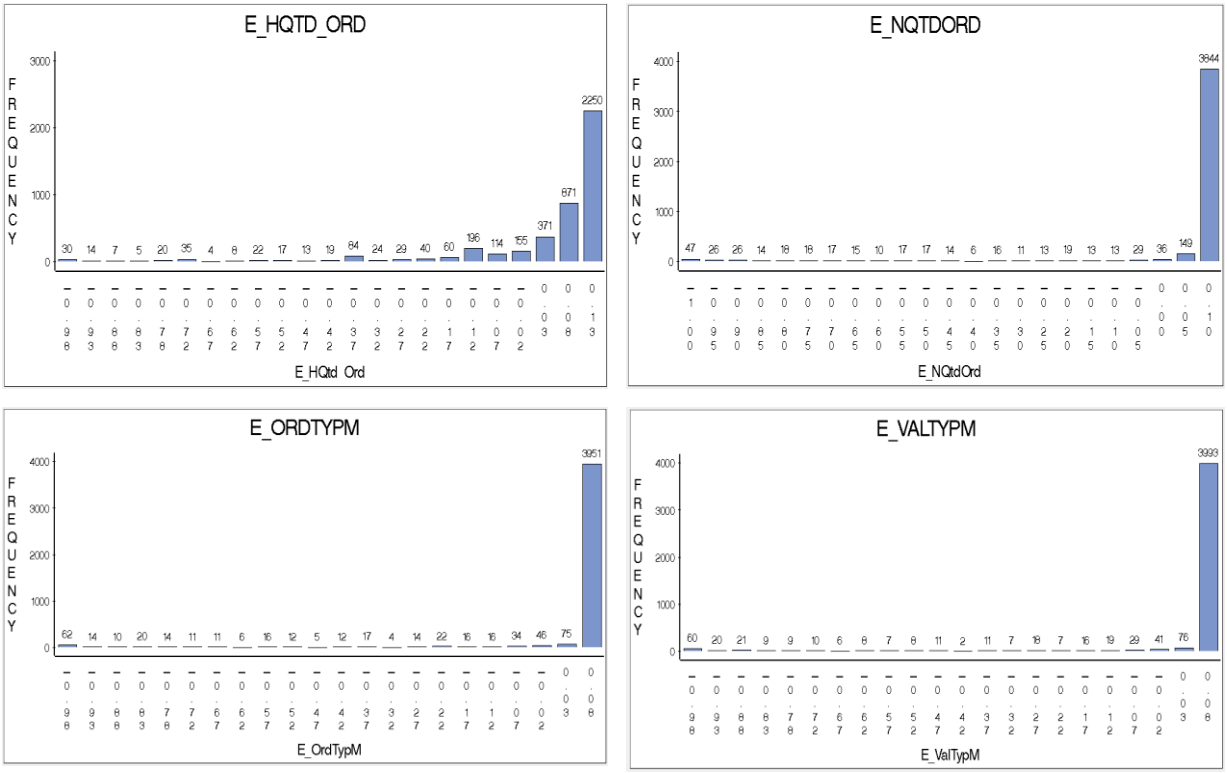
Variáveis	Descrição	Formato
Dat_Ses	Data da Sessão de bolsa	DDMMAAAA
ISIN	Código de identificação do instrumento	Char (12)
Cod_If	Código de identificação do Intermediário Financeiro	Num (8)
Ord_Stat	Estado da oferta inserida na sessão de bolsa: 0-New; 1-Partially filled; 2-Filled 3-Done for Day; 4-Cancelled; 5-Replaced; 8-Rejected; C-Expired; S-Cancelled by Market Operation; 0-Eliminated by Coraporate Event.	Char (1)
Ord_Time	Hora de registo da oferta no sistema de negociação.	(HHMMSS999999)
Cnc_Dat	Hora de cancelamento/modificação/execução da oferta.	(HHMMSS999999)
Ord_Ori	Origem da oferta: 1-Client, 2-House 6-Liquidity Provider, 7-Related Party, 8-Riskless Principal.	Num(1)
Val_Typ	Prazo de validade da oferta: 0-Day, 1-GTC, 2-VFA, 3-IOC, 4-FOK, 6-GTD, 7-VFC; V- good for VWAP.	Char (1)
Ord_Typ	Tipo de oferta: 1-Market, 2-Limit, 3-Stop, 4-StopLimit, P-Pegged, K-Market to limit	Char (1)
Qtd_Ord	Quantidade de títulos registados para oferta.	Num (12)
Pre_Ofer	Preço da oferta	Num(12,4)
Ord_id	Número da oferta identificada pelo sistema de negociação.	Num (8)

Quadro 3.2 – *Base de dados de Negócios -Variáveis Originais (Detalhe)*

Variáveis	Descrição	Formato
Dat_Ses	Data da Sessão de bolsa	DDMMAAAA
Cod_ISIN	Código de identificação do instrumento	Char (12)
Num_Not	Número do negócio	Num (8)
Hor_Neg	Hora do negócio	(HHMMSS)
Pre_Neg	Preço do negócio	Num (12,4)
Qtd_Neg	Quantidade do negócio	Num (15)
Dat_Ordc	Data da oferta de compra	DDMMAAAA
Dat_Ordv	Data da oferta de venda	DDMMAAAA
Num_Ordc	Número da oferta de compra	Num (8)
Num_Ordv	Número da oferta de venda	Num (8)

7.3. Distribuição das variáveis (normalizadas) utilizadas no treino do SOM – Instrumento financeiro "A".





7.3.1. Valor Mínimo, Máximo e Médios (normalizados) das variáveis utilizadas para o treino do SOM – instrumento financeiro "A"

Name	Min	Max	Mean	Std Dev.	Missing %	Skewness	Kurtosis
DAT_IF	406952	4.08E7	5.83E6	8.65E6	0%	3.7232	12.18
ORD_DAT	18779	18900	18840	35.469	0%	-0.012	-1.176
NUM_IF	2	102	50.924	30.592	0%	0.1381	-1.223
A_RACIOORD	-1	0.1668	0.048	0.2348	0%	-3.284	10.235
A_ORDEXECM	-1	0.0532	0.0284	0.1072	0%	-7.618	63.664
B_ORDMODM	-1	0.0868	0.0416	0.1835	0%	-4.556	20.184
C_ORDTIMEM	-1	0.0918	0.0473	0.1839	0%	-4.795	22.341
D_ORDORIM	-1	0.0861	0.0422	0.1834	0%	-4.625	20.672
E_ORDTYPM	-1	0.0936	0.0424	0.1903	0%	-4.311	18.034
E_VALTYPM	-1	0.0911	0.0445	0.1853	0%	-4.604	20.569
E_HQTD_ORD	-1	0.1469	0.032	0.1984	0%	-2.982	9.5975
E_NQTDORD	-1	0.1051	0.0434	0.2087	0%	-3.861	13.98

7.4. Grupos de variáveis utilizadas no treino do SOM.

7.4.1 – Variáveis utilizada no treino da 1º etapa.

Name	Model Role
DAT_IF	id
ORD_DAT	rejected
COD_INST	rejected
NUM_IF	rejected
A_RACIOORD	input
A_ORDEXEC	input
B_ORDMOD	input
C_ORDTIME	input
D_ORDORI	input
E_ORDTYP	input
E_VALTYP	input
E_HQTD_ORD	input
E_NQTDORD	input

Name	Min	Max	Mean	Std Dev	Missing	Skewness	Kurtosis
A_ORDEXEC	-0.999	0.257	0.0504	0.2977	0%	-1.962	3.3739
A_RACIOORD	-1	0.1668	0.0488	0.2347	0%	-3.32	10.526
B_ORDMOD	-0.999	0.1971	0.0496	0.2979	0%	-2.158	3.5604
C_ORDTIME	-1	0.0918	0.0457	0.188	0%	-4.782	22.173
DAT_IF	406952	4.08E7	5.83E6	8.65E6	0%	3.724	12.192
D_ORDORI	-0.998	0.1957	0.0531	0.2733	0%	-2.365	5.1378
E_HQTD_ORD	-1	0.1469	0.0362	0.1904	0%	-3.014	9.9898
E_NQTDORD	-1	0.1051	0.0438	0.2035	0%	-3.802	13.64
E_ORDTYP	-0.996	0.2853	0.0488	0.3188	0%	-1.742	2.3118
E_VALTYP	-0.998	0.2671	0.0488	0.3166	0%	-1.752	2.334
NUM_IF	2	102	51.048	30.44	0%	0.1285	-1.208
ORD_DAT	40695	40816	40756	35.219	0%	-0.03	-1.183

7.4.2 – Variáveis utilizada no treino da 2º etapa.

Name	Model Role
DAT_IF	id
ORD_DAT	rejected
COD_INST	rejected
NUM_IF	rejected
A_RACIOORD	input
A_ORDEXECM	input
B_ORDMODM	input
C_ORDTIMEM	input
D_ORDORIM	input
E_ORDTYPM	input
E_VALTYPM	input
E_HQTD_ORD	input
E_NQTDORD	input

Name	Min	Max	Mean	Std Dev.	Missing %	Skewness	Kurtosis
DAT_IF	406952	4.08E7	5.83E6	8.65E6	0%	3.7232	12.18
ORD_DAT	18779	18900	18840	35.469	0%	-0.012	-1.176
NUM_IF	2	102	50.924	30.592	0%	0.1381	-1.223
A_RACIOORD	-1	0.1668	0.048	0.2348	0%	-3.284	10.235
A_ORDEXECM	-1	0.0532	0.0284	0.1072	0%	-7.618	63.664
B_ORDMODM	-1	0.0868	0.0416	0.1835	0%	-4.556	20.184
C_ORDTIMEM	-1	0.0918	0.0473	0.1839	0%	-4.795	22.341
D_ORDORIM	-1	0.0861	0.0422	0.1834	0%	-4.625	20.672
E_ORDTYPM	-1	0.0936	0.0424	0.1903	0%	-4.311	18.034
E_VALTYPM	-1	0.0911	0.0445	0.1853	0%	-4.604	20.569
E_HQTD_ORD	-1	0.1469	0.032	0.1984	0%	-2.982	9.5975
E_NQTDORD	-1	0.1051	0.0434	0.2087	0%	-3.861	13.98

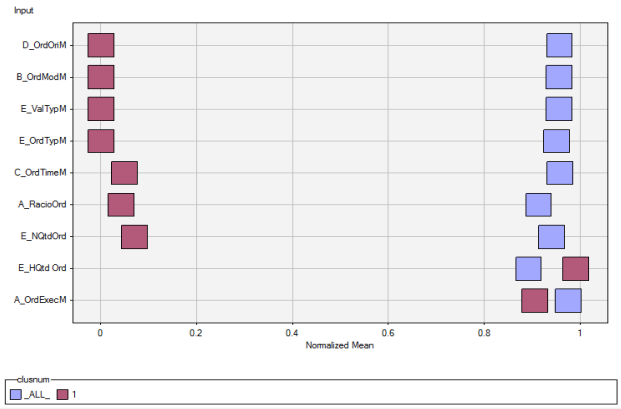
7.4.3 – Variáveis utilizada no treino da 3º etapa.

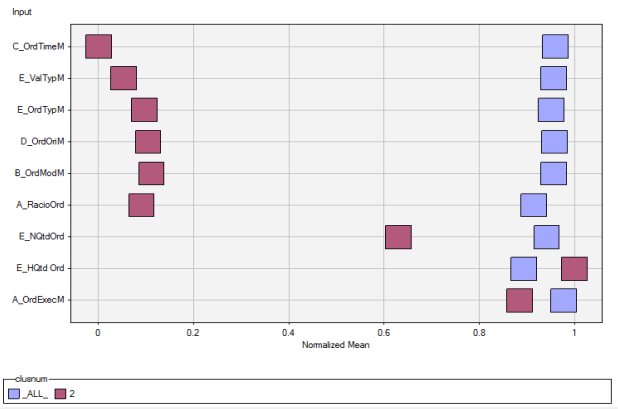
Name	Model Role
DAT_IF	id
ORD_DAT	rejected
COD_INST	rejected
NUM_IF	rejected
A_RACIOORD	input
A_ORDEXEC	input
B_ORDMOD	input
C_ORDTIME	input
D_ORDORI	input
E_ORDTYPM	input
E_VALTYPM	input
E_HQTD_ORD	input
E_NQTDORD	input

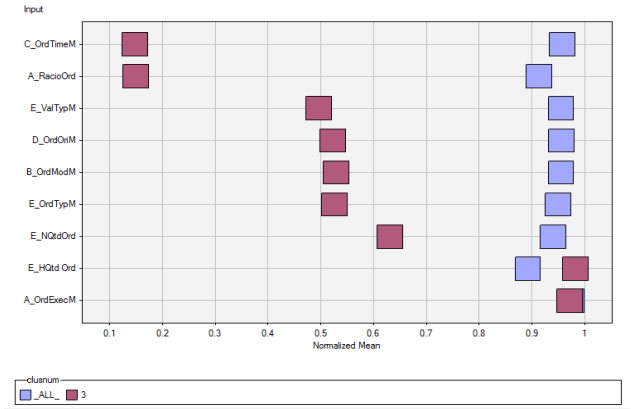
Name	Min	Max	Mean	Std Dev.	Missing %	Skewness	Kurtosis
DAT_IF	406952	4.08E7	5.85E6	8.69E6	0%	3.7048	12.046
ORD_DAT	18779	18900	18840	35.876	0%	-0.015	-1.207
NUM_IF	2	102	50.924	30.716	0%	0.1482	-1.233
A_RACIOORD	-1	0.1668	0.0488	0.2369	0%	-3.297	10.191
A_ORDEXEC	-0.999	0.257	0.0439	0.3041	0%	-1.878	2.9363
B_ORDMOD	-0.999	0.1971	0.0534	0.2936	0%	-2.191	3.7178
C_ORDTIME	-1	0.0918	0.0445	0.1918	0%	-4.656	20.825
D_ORDORI	-0.998	0.1957	0.047	0.2834	0%	-2.25	4.3934
E_ORDTYPM	-1	0.0936	0.0415	0.1937	0%	-4.257	17.467
E_VALTYPM	-1	0.0911	0.0416	0.193	0%	-4.414	18.654
E_HQTD_ORD	-1	0.1469	0.0351	0.194	0%	-2.994	9.76
E_NQTDORD	-1	0.1051	0.0465	0.2037	0%	-4.037	15.474

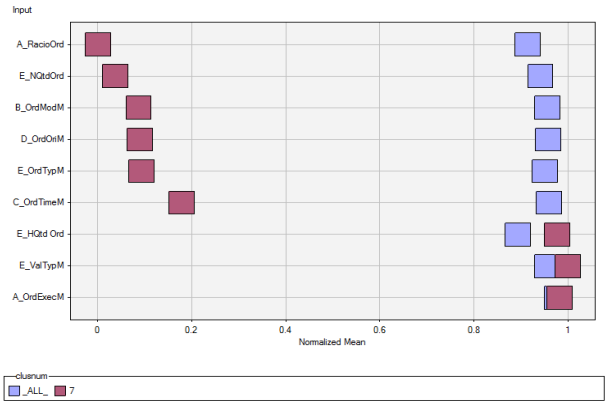
7.5. Mapa topológico (4*6) - opção *default* do SAS Entreprise Miner

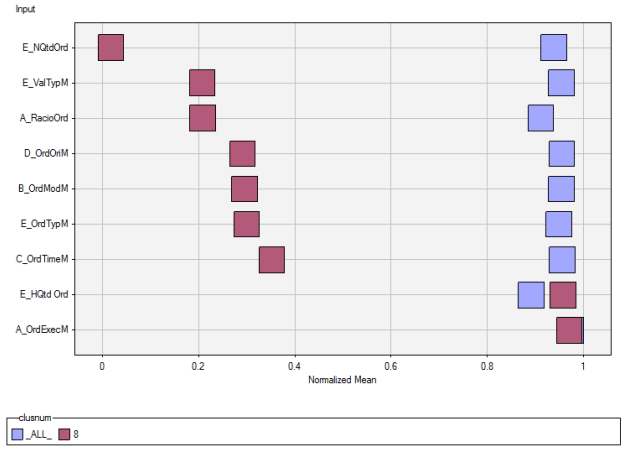
As figuras *infra* identificam o exemplo de sete dos 24 *clusters* obtidos com o método *Batch* SOM, com grau de dissemelhança pretendidos de serem encontradas com o treino da rede SOM. (*output do SAS Enterprise Miner*).

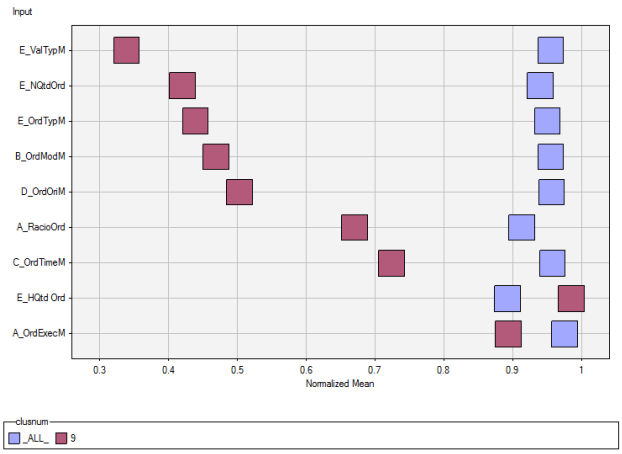


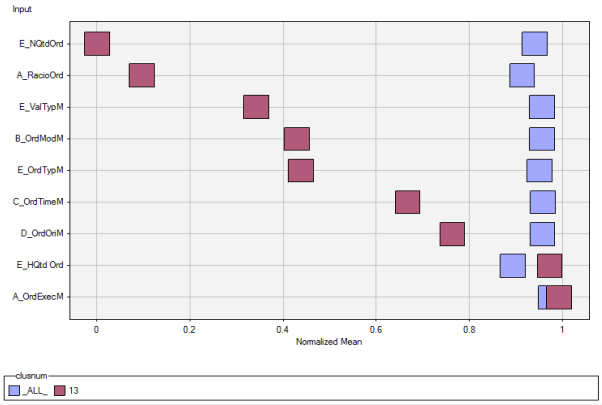



Input
C_OrdTimeM
A_RacioOrd
E_ValTypM
D_OrdOrtM
B_OrdModM
E_OrdTypM
E_NQtdOrd
E_HQtd Ord
A_OrdExecM
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Normalized Mean
clusnum
ALL
3


Input
A_RacioOrd
E_NQtdOrd
B_OrdModM
D_OrdOrtM
E_OrdTypM
C_OrdTimeM
E_HQtd Ord
E_ValTypM
A_OrdExecM
0
0.2
0.4
0.6
0.8
1
Normalized Mean
clusnum
ALL
7


Input
E_NQtdOrd
E_ValTypM
A_RacioOrd
D_OrdOrtM
B_OrdModM
E_OrdTypM
C_OrdTimeM
E_HQtd Ord
A_OrdExecM
0
0.2
0.4
0.6
0.8
1
Normalized Mean
clusnum
ALL
8


Input
E_ValTypM
E_NQtdOrd
E_OrdTypM
B_OrdModM
D_OrdOrtM
A_RacioOrd
C_OrdTimeM
E_HQtd Ord
A_OrdExecM
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Normalized Mean
clusnum
ALL
9


Input
E_NQtdOrd
A_RacioOrd
E_ValTypM
B_OrdModM
E_OrdTypM
C_OrdTimeM
D_OrdOrtM
E_HQtd Ord
A_OrdExecM
0
0.2
0.4
0.6
0.8
1
Normalized Mean
clusnum
ALL
13

7.5.1. Matriz de distâncias resultante do mapa topológico (4*6) - opção *default* do SAS Entreprise Miner.

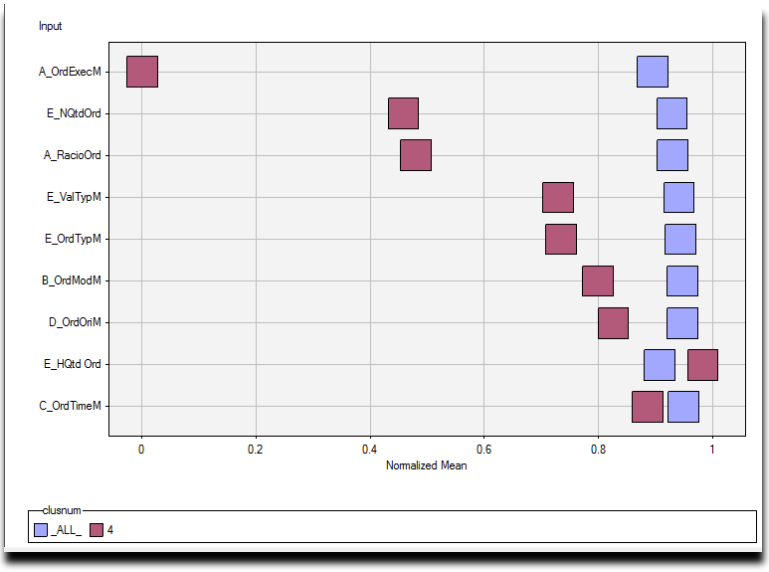
A figura *infra* (*output do SAS Enterprise Miner*) identifica a matriz de distância obtida com o treino de SOM para um mapa topológico de 4 linhas e 6 colunas. O métdo *Batch* SOM, classifica para algumas varíaveis indivíduos com grau de dissemelhança significativo face a população (como verificado nas figuras supra) mas com *cluster* bastantes próximos, o que exige a uma redefinição do número de segmentos (*clusters*).

SEGMNT	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster
7	29	0.1438605822	0.710375298	8	0.9707003287
3	8	0.1278777879	0.5705789552	8	0.8390266545
4	14	0.2140144323	0.8158810252	10	0.7376935675
5	3	0.0920864995	0.3000164594	6	0.7018531986
10	21	0.1316388059	0.9413871616	16	0.6773230199
13	22	0.1400083041	0.6871718656	8	0.6587356801
8	34	0.123418519	0.7723207499	12	0.6587356801
9	8	0.1399519341	0.6331776674	14	0.6286489295
2	27	0.1112167171	0.8951751561	1	0.6020360642
1	62	0.1022993281	0.9653064095	2	0.6020360642
14	18	0.1343183854	0.5703752929	19	0.5548715433
19	21	0.1321683963	0.655143693	14	0.5548715433
21	39	0.0833654702	0.5152571408	22	0.5496722298
20	55	0.0783579954	0.4512161463	15	0.455092896
15	34	0.0918204601	0.5991531984	22	0.4509689053
6	138	0.0520372038	0.5363390161	12	0.4203069556
16	41	0.0605986746	0.3158370325	11	0.3198861764
11	9	0.0738644526	0.3018628188	16	0.3198861764
12	200	0.0303573857	0.1930168017	17	0.2975325043
22	88	0.0579082715	0.3641355289	23	0.2541795184
17	48	0.0389893593	0.2641944514	18	0.220899362
18	559	0.0231582576	0.2014745957	22	0.1847318086
23	710	0.0302638564	0.9847866144	24	0.0996439349
24	2200	0.0187148484	0.6767452042	23	0.0996439349

7.6. Mapa topológico (2*3) – Outros *outputs* do SAS Entreprise Miner

7.6.1 – Detalhe do *cluster* 4, obtido no treino do SOM na 4º etapa.

A figura *infra* identifica os valores médios do *cluster* 4 face a média da população. A par de apresentar, para a mairia das variáveis valores dissemelhantes da população, objectivo do estudo, a variável que mais relevância apresenta para formação do *cluster* (“A_OrdExecM”) é ao mesmo tempo a mais dissemenlhantes dentre as demais variáveis. No entanto, para obter o resultado do estudo, espera-se que esta variável apresente o maior grau de semelhança possível com a população. Portanto este *cluster* não é relevante para o estudo em questão.



7.7. Detalhe da classificação dos individuos aos *clusters*, resultante dos mapas topológicos utilizados no treino do SOM.

7.7.1. *Cluster* 1 – Cluster seleccionado no processo e treino da rede.

CLUSTER 1

Intermediários Financeiros						
Clusters	**13**	**15**	**43**	**66**	**87**	**101**
1		1		39		22
2				27		
3				4		3
7				14		15
8	2	1				31
9	2					
13			1		7	7
Total	**4**	**2**	**1**	**84**	**7**	**78**

Clusters de origem (mapa topológico 4*6): 1,2,3,7,8,9 e 13

7.7.2. *Cluster* 2 – Maior Frequência.

Este *cluster* contém a maior frequência de registos (3.002) face aos demais 5 clusters e apresenta para todas as variáveis valores médios superiores (efeito sigmoidal) a média da população.

CLUSTER 2

Intermediários Financeiros

Clusters	**2**	**3**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**13**	**14**	**15**	**16**	**17**	**19**	**20**	**21**	**24**	**26**	**29**	**31**	**32**	**33**	**34**	**37**	**39**	**40**	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**
10															1			5																	
15										2																									
16								1		4					1			16		6															
18						1					1		2						1					1							1		1		
22	10			3						10	1	1													4		2	4	1	3	10		1		
23	30			17				4		5	6				6		3	2	4	56	1	38			5		45	30	14		6		11		5
24	33	2	4	47	50	39	54	55	47	1	28	6	19	43	47	71	8	58	55	8	14	22	36	61	60	3	26	15	34	11	44	29	35	60	49
Total	**73**	**2**	**4**	**67**	**50**	**40**	**54**	**60**	**47**	**22**	**36**	**7**	**21**	**43**	**55**	**71**	**11**	**81**	**60**	**70**	**15**	**60**	**36**	**62**	**69**	**3**	**73**	**49**	**49**	**14**	**61**	**29**	**48**	**60**	**54**

Clusters	**49**	**50**	**51**	**53**	**55**	**56**	**57**	**58**	**63**	**64**	**65**	**66**	**67**	**68**	**69**	**70**	**73**	**75**	**76**	**77**	**84**	**86**	**87**	**88**	**89**	**90**	**91**	**92**	**93**	**94**	**96**	**97**	**99**	**100**	**102**
10																											1								
15																																			
16			2											4														4						1	2
18						2								2																					
22		1		1					1				2													1	1							10	
23		4	54	3	1	1			31		2		9	17		3			19	8		39				8	41	56	8	3	1	2	8	46	21
24	1	28	19	41	4	47	5	1	11	80	53	1	48	44	14	44	1	70	36	17	2	10	14	31	63	25	24	15	64	4	48	58	25	24	54
Total	**1**	**33**	**75**	**45**	**5**	**50**	**5**	**1**	**43**	**80**	**55**	**1**	**59**	**67**	**14**	**47**	**1**	**70**	**55**	**25**	**2**	**49**	**14**	**31**	**63**	**34**	**67**	**75**	**72**	**7**	**49**	**60**	**33**	**81**	**77**

Clusters de origem (mapa topológico 4*6): 10,15,16,18,22,23 e 24.

7.7.3. *Cluster* 3 – Grande quantidades

Cluster com a menor frequência de registos (142) face aos demais 5 clusters. Apresenta destaque para a variável "E_HqtdOrd", única variável com valores médios distintos da média da população dos *clusters*. A referida variável identifica moda das quantidades registadas nas ofertas pelos indivíduos do *cluster*, sendo esta (quantidade) elevada. As demais variávies apresentam valores próximos da média da população.

CLUSTER 3

Intermediários Financeiros																															
Clusters	**5**	**7**	**8**	**10**	**11**	**14**	**16**	**19**	**20**	**26**	**31**	**32**	**39**	**40**	**45**	**46**	**47**	**48**	**55**	**56**	**63**	**65**	**69**	**84**	**86**	**89**	**90**	**91**	**93**	**94**	**96**	**99**	**100**
5															3																		
6	2	4	1	3	1	3	2	1	1	5	16	2	1	4	38	1	6	1	2	2	1	3	3	1	3	2	3	3	2	3	12	5	1
10																		1															
Total	**2**	**4**	**1**	**3**	**1**	**3**	**2**	**1**	**1**	**5**	**16**	**2**	**1**	**4**	**41**	**1**	**6**	**2**	**2**	**2**	**1**	**3**	**3**	**1**	**3**	**2**	**3**	**3**	**2**	**3**	**12**	**5**	**1**

Clusters de origem (mapa topológico 4*6): 5,6 e 10.

7.7.4. *Cluster* 4 – Elevadas execuções de ofertas

Este *cluster* contém a terceira menor frequência de registos (207), sendo caracterizado pela variável "A_OrdExecM ", variável que identifica que os indivíduos (IF) seleccionados para o cluster têm um elevado nível de execução das suas ofertas, bem como uma frequência significativa de ofertas com quantidades equivalentes a moda (Variável E_NqtdOrd).

CLUSTER 4

Intermediários Financeiros																
Clusters	**2**	**7**	**13**	**15**	**19**	**20**	**24**	**29**	**37**	**43**	**44**	**63**	**66**	**86**	**87**	**101**
3													1			
4		1	8		1		1	1				1		1		
9			5												1	
10			2			1	2							1		
13									1						3	3
14			13						1							4
15			26	3						1	1					
19										12					9	
20			3							50					2	
21	3								12		19	4				
22	1		3						1	2	3					
Total	**4**	**1**	**60**	**3**	**1**	**1**	**3**	**1**	**15**	**65**	**23**	**5**	**1**	**2**	**15**	**7**

Clusters de origem (mapa topológico 4*6): 3,4,9,10,13,14,15,19,20,21 e 22.

7.7.5. Cluster 5 – Quantidades significativas

Este *cluster* contém a quarta menor frequência de registos (211), sendo caracterizado, a semelhança do *cluster* 3, pela variável "E_HQtdOrd", com valores médios superiores a média da população. A moda das quantidades registadas nas ofertas pelos indivíduos do *cluster*, sendo significativa é inferior a verificada no *cluster* 3.As demais variávies apresentam valores próximos da média da população.

CLUSTER 5

Intermediários Financeiros

Clusters	2	3	5	6	7	8	10	11	14	16	19	20	21	26	29	31	32	34	40	41	42	45
10											1											
11		1								1	1											
12	4		1	2	5	6	4	9	1	19	3	3	1	10	1	13	3	3	7	1	1	15
17	1																					
18																						
21																						
Total	**5**	**1**	**1**	**2**	**5**	**6**	**4**	**9**	**1**	**20**	**5**	**3**	**1**	**10**	**1**	**13**	**3**	**3**	**7**	**1**	**1**	**15**

Clusters	46	47	48	49	51	55	56	57	63	64	65	67	68	69	76	86	89	90	91	94	96	99
10																						
11																						
12	11	7	1	2	2	1	2	1	1	1	9	1	1	12	2	2	6	8	3	4	3	8
17									3							1						
18		1																				
21									1													
Total	**11**	**8**	**1**	**2**	**2**	**1**	**2**	**1**	**5**	**1**	**9**	**1**	**1**	**12**	**2**	**3**	**6**	**8**	**3**	**4**	**3**	**8**

Clusters de origem (mapa topológico 4*6): 10,11,12,17,18 e 21.

7.7.6. *Cluster* 6 – Quantidades relativas

Este cluster contém cerca de 650 registos, sendo caracterizado, a semelhança dos *cluster* 3 e 5, pela variável "E_HQtdOrd", no entanto as demais variáveis apresentam valores diferentes e não tão próximos da população. Verifica-se, pelos *clusters* de origem (mapa topológico 4*6) que parte a actuação de alguns individuos esta subdividade entre os clusters 5 e 6, o que poderia induzir a uma redução do número de clusters. No entanto, os *clusters* de interesse – os que contribuem para o agrupamento formado pelo cluster 1 (clusters 1,2,3,7,8,9 e 13) não encontram-se entre os clusters 5 e 6, o que se supõe não trará qualquer acréscimo a estrutura então seleccionada.

CLUSTER 6

Intermediários Financeiros

Clusters	2	5	6	7	8	9	10	11	14	15	16	19	20	21	24	26	29	31	32	33	34	39	40	42	44	46	47
10					1						1	2															
11				1											1		2										
15																											
17		1															2										
18	2		7	24	37	3	19	18	30	1	42	22	10	1		11	6	7	18	3	10	1	1	1		23	12
22	1		1																						1		
23					1												4		3						2		
Total	**3**	**1**	**8**	**25**	**39**	**3**	**19**	**18**	**30**	**1**	**43**	**24**	**10**	**1**	**1**	**11**	**14**	**7**	**21**	**3**	**10**	**1**	**1**	**1**	**1**	**25**	**12**

Clusters	48	50	51	53	56	63	64	65	67	68	69	75	76	86	88	89	90	91	92	93	94	96	97	99	100	102
10			1														1									
11						2																				
15														1												
17		4		1		24			2				2	3				3		1						
18	3	41	3		4	1	1	19	9	16	21	1	24	13	1	14	16	6	9	4	3	4	1	19	2	2
22						5								1					1						1	
23	3		4							2			1	10			1	3	1					1	1	
Total	**6**	**45**	**8**	**1**	**4**	**32**	**1**	**19**	**11**	**18**	**21**	**1**	**27**	**28**	**1**	**14**	**18**	**12**	**11**	**5**	**3**	**4**	**1**	**20**	**4**	**2**

Clusters de origem (mapa topológico 4*6): 10,11,15,17,18, 22 e 23.

7.8. Detalhe da classificação dos individuos aos *clusters*, resultante da validação do SOM.

CLUSTER: 1 **Acções "M"**

Nº sessões: 76

Intermediário Financeiro

Cluster Origem	13	42	44	51	63	101	Total
1	26		1			38	65
2	28		4			1	33
3	6		6			1	13
4		1		1			2
8			1			17	18
9	1		4		1	6	12
13						1	1
Total	**61**	**1**	**16**	**1**	**1**	**64**	**144**
% Part.	80,3%	1,3%	21,1%	1,3%	1,3%	84,2%	

CLUSTER: 1 **Acções "J"**

Nº sessões: 85

	Intermediário Financeiro						
Cluster Origem	**13**	**44**	**46**	**63**	**66**	**86**	**Total**
1	51						51
2	17	1	1				19
3	2						2
4					1		1
8	6	4		3		1	14
9	4	4		1			9
13		1					1
Total	**80**	**10**	**1**	**4**	**1**	**1**	**97**
% Part.	94,1%	11,8%	1,2%	4,7%	1,2%	1,2%	

CLUSTE 1 **Acções "I"**

Nº sessões: 86

	Intermediário Financeiro			
Cluster Origem	**2**	**13**	**44**	**Total**
1		47		47
2		16		16
3	1	4	1	6
8		3		3
9	1	4	6	11
Total	**2**	**74**	**7**	**83**
% Part.	2,3%	86,0%	8,1%	

CLUSTE 1 **Acções "E"**

Nº sessões: 85

	Intermediário Financeiro				
Cluster Origem	**13**	**44**	**55**	**101**	**Total**
1	57	2		1	60
2	22	2			24
3	2		1		3
8	1	3		1	5
9		2			2
13		2			2
Total	**82**	**11**	**1**	**2**	**96**
% Part.	96,5%	12,9%	1,2%	2,4%	

CLUSTER: 1 **Acções "B"**

N° sessões: 85

Intermediários Financeiros

Cluster Origem	13	24	37	44	56	101	Total
1	6					29	35
2	26					2	28
3	16					11	27
4	1						1
7	13					14	27
8						20	20
9	3	1		1	2		7
13			1				1
Total	**65**	**1**	**1**	**1**	**2**	**76**	**146**
% Part.	76,5%	1,2%	1,2%	1,2%	2,4%	89,4%	